Alma Mater Studiorum
Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Modelling complex systems in the severely undersampled regime: a Bayesian Model Selection approach

Relatore:
Prof. Armando Bazzani

Correlatore:
Prof. Matteo Marsili

Presentata da:
Rocco Mantovani

Anno Accademico 2017/2018

Abstract

L'inferenza di modelli di spin è uno strumento diffuso nell'approccio statistico ai sistemi complessi. Tipicamente ci si limita a modelli con interazioni a uno e due corpi: per il principio di massima entropia, ciò equivale ad assumere che magnetizzazioni e correlazioni a due unità costituiscano le variabili rilevanti (statistiche sufficienti) del sistema. L'assunzione non è giustificata nel caso generale; il problema della selezione tra modelli con interazioni di ordine arbitrario è però alto-dimensionale. Esso può essere affrontato tramite una particolare euristica Bayesiana che permette di ottenere le variabili rilevanti direttamente dal campione; la selezione avviene nella classe delle *misture*, e i risultati vengono proiettati sulla rappresentazione di spin. Il risultato è l'ottenimento delle statistiche sufficienti senza alcuna assunzione a priori. Il numero di tali statistiche è modulato da quello di differenti frequenze empiriche nel campione; in regime di sottocampionamento, esso è molto minore della dimensione del modello completo. Ciò rende il problema di inferenza dei parametri tipicamente basso-dimensionale. Il principale scopo di questo lavoro è quello di investigare esplicitamente come l'informazione sia organizzata nella mappa tra misture e modelli di spin. La comprensione dettagliata di tale mappa suggerisce nuovi approcci per la regolarizzazione; inoltre i risultati gettano luce sulla natura delle statistiche sufficienti, che risultano essere funzioni degli stati solo tramite le frequenze empiriche di questi. Mostriamo come da un approccio integralmente Bayesiano emerga sotto opportune condizioni un termine regolarizzatore "L2"; verifichiamo numericamente se tali condizioni sono tipicamente soddisfatte. Presentiamo infine alcune osservazioni qualitative circa l'emersione di *loop stuctures* nella mappa da misture a spin; queste aprono scenari interessanti per la ricerca futura.

Contents

Introduction

> A chemist notices a surprising phenomenon. Now if he has a high admiration of Mill's Logic,... he must work on the principle that, under precisely the same circumstances, like phenomena are produced. Why does he then not note that this phenomenon was produced on such a day of the week, the planets presenting a certain configuration, his daughter having on a blue dress, he having dreamed of a white horse the night before, the milkman having been late that morning, and so on? The answer will be that in early days chemists did use to attend to some such circumstances, but that they have learned better.
>
> Charles Sanders Peirce [1]

Modelling: a verbous introduction

Abstract modelling, as the ability of building an inner model of the environment and using it to make predictions about that environment, is something which is not at all peculiar of human beings. Any living creature's success in the two fundamental - and universally shared - tasks of survival and reproduction is driven by the quality of their choices in relation to the environment they live in. These choices arise, directly or indirectly, from models - and the very notion of "choice" here does not need to raise issues regarding the presence or absence of something that we could call "consciousness": we can (sloppily - but effectively) *define* choices to address emergences of functional, nontrivial behaviors in living systems - from the policies and strategies adopted by ant colonies in the complicated task of food foraging (strategies that must necessarily be modulated by contingent observations of the environment's states), to more "bottom" behaviors like the switching between metabolic states in elementary organisms (modulated by the "observed" concentrations of nutrients and toxins in the surroundings).
This elementary capability of making *predictions* (be them implicit or explicit) out of observations of the environment, and "choosing" effective policies relying on these, is then enough to characterize what we mean when we speak of "modelling" at the most abstract level possible.

Scientific modelling comes into play the moment we let mathematics in, and allow it to be used as the main tool and language at our disposal for developing descriptive schemes of the world, and extracting preditions out of them. The very definition of what the boundaries of science are exhibits implicitly the strict requirement of a *mathematical* kind of reasoning: if this wasn't the case, we would probably include into the castle of science some astonishing medieval "outliers" like Johannes Buridanum (XIV century - father of the theory of *impetus*), Thomas Bradwardine, and Nicole d'Oresme, all of which developed, with different fortune, pre-mathematical versions of quantities that later became part of the foundations of classical physics (see, for instance, [2] and references within).
Even before the choice of a logical scheme, science by constitution requires exploitation of the *unreasonable effectiveness of mathematics* [3] in describing natural phenomena. This can be done, and has been done, both in a more top-down, "empirical", inductive, statistically oriented fashion (e.g. empirical laws in early thermodynamics) and a bottom-up, "ab initio", deductively oriented one (e.g. statistical mechanics). Science has developed as a complicated intertwining of instances with these different directions; yet the *objectives* have always been shared, and so the fundamental practices.

In recent times - with the Information Technology revolution, the sudden availability of huge amounts of data and the consequent change of focus towards the so-called *complex systems* - the distinction between *physical* and *statistical* modelling has become more and more important.
A *physical* model is tipically a product of a long logical process, in alternating stages of induction, abduction, and deduction, leading ultimately to a theory which yiels a representation of the experimental results which is the *simplest* possible. A lot of energy is spent in the task of identifying which of the many available observables are *relevant* and which are not - this being a fact which is often not fully recognized. The quote from C.S.Peirce at the beginning of this chapter expresses this precise idea; the same observations can be found in Wigner [3]:
"It is, as Schrodinger has remarked, a miracle that in spite of the baffling complexity of the world, certain regularities in the events could be discovered. One such regularity, discovered by Galileo, is that two rocks, dropped at the same time from the same height, reach the ground at the same time. [...] the regularity which we are discussing is independent of so many conditions which could have an effect on it. It is valid no matter whether it rains or not, whether the experiment is carried out in a room or from the Leaning Tower, no matter whether the person who drops the rocks is a man or a woman. It is valid even if the two rocks are dropped, simultaneously and from the same height, by two different people. [...] It is the skill and ingenuity of the experimenter which show him phenomena which depend on a relatively narrow set of relatively easily realizable and reproducible conditions."
The process of selecting relevant variables ensures that the physical model

will depend on as few parameters as possible; once this model is established, it will yield correct preditions because it will capture essential features of the phenomenon that it describes - and will be obviously specific for that specific phenomenon.

The situation is completely different when we look at *statistical* models; quoting from [4]:
"[...] The Big Data deluge has shown that understanding is no more necessary to solve problems, such as image or speech recognition and language translation. A statistical model trained on a sufficiently large number of instances of the solution can learn how to generalize from examples in ways that resemble the "unknown rules" that humans follow".
This holds inasmuch as the notion of "solving" we look at is the one of "making correct and useful predictions" - this notion being, after all, a pretty reasonable one! Yet, if our task is to uncover possible fundamental laws and principles governing the system under study, it is highly probable that such models won't provide the best possible weaponry to do that; this mainly because of their dependence on a huge number of parameters, and their nonspecificity (for instance, a DNN with one and the same architecture can be trained to solve many different problems), qualities that place these models very far from the *simplicity* we desire for insightful descriptions.

This thesis

This thesis is concerned with *statistical model selection* as a tool to enforce a request of simplicity while addressing some specific classes of inverse problems in a statistical mechanical framework. We rely mainly on [5], describing the method devised in it, correcting some minor errors, and adding some results and insights, both from an analytical and a numerical point of view.

- Chapter 1 will be devoted to a comprehensive introduction to statistical model selection. We will first discuss how we can regard statistical parametric models as manifolds in the space of probability distributions, and how we can use these models' parameters as coordinates on these manifolds. We'll then see how the notion of *distinguishability* of distributions induces a natural measure on these manifolds; we will discuss how such measure can be normalized to become a sound *prior* for the Bayesian learning of parameters. Lastly, we will define a quantity called *geometric complexity* of a model, and argue that we can interpret it as a measure of its "simplicity".

- In Chapter 2 basic notions about spin models with interactions of arbitrary order are presented, and a Bayesian-grounded procedure of model selection is attempted in the general case within this class; this

will bring up a geometric complexity term, of the type introduced in the preceding chapter; discussion of this term will lead us to the conclusion that uninformative Bayesian analysis alone *does not justify* restriction to at-most-pairwise interactions between spins. We close by noting that the number of spin models with interaction of arbitrary order is huge, and selection performed directly within this class in is practice unfeasible.

The first two chapters can be thought as a long introduction, to both understand with conceptual motivation for the subsequent exposition, and familiarize with the notation. The main point one has to get for all what follows is the *absence of theoretical justification* for restriction to low-order spin models from an information-theoretic point of view.

- In Chapter 3 we discuss the method devised in [6] to actually obtain an effective selection within generalized spin models in a feasible manner: this will be done by first performing model selection on the class of *mixture models*, where this task turns out to be straightforward, and then projecting the results on the spin representation, via a linear application χ mapping mixture parameters into spin ones, and containing all the information about the selected mixture model. This will lead to the identification of a small group of *sufficient statistics* in the form of specific linear combinations of spin operators. We discuss regularization and stability issues for the parameter estimation stage on selected models. We then close presenting some numerical results.

- In Chapter 4 we give a full, in detail characterization of how information about different sets of the mixture is organized inside the matrix χ. We obtain an analytical characterization of the spectrum of this matrix, whose interpretation possibly leads to a previously unconsidered method of regularization. As a bonus, we present a qualitative finding linking mixture models with families of spin *loops*; this section is not formal, and serves merely as motivation for future investigations.

- In Chapter 5 we complicate the framework used in Chapter 3 by adopting a fully Bayesian approach for mixture selection, and try to see how the act of averaging over models' posterior probabilities affects the definition of our sufficient statistics. We find that, given that some reasonable conditions are met, we can analitically derive from this approach a loss function for parametric estimation containing a "L2" regularizing term. We then run simulations to check if the necessary conditions are met in typical cases, and discuss the results.

Original contributions

Chapter 3 presents already existing work; we add some original interpretations of results, an additional numerical result, and correct a few former imprecisions. Chapters 4 and 5 are entirely composed of original work by the authors.

Chapter 1

Model Selection in statistical inference

Introduction: modelling complex systems

Big data is not enough. The recent revolution in our capabilities of extracting huge amounts of observational data in a wide class of systems, from biological to financial ones, has surely permitted an outstanding jump forward in a plethora of applications, but has also raised an important issue: while extracting data has become easier by orders of magnitude, extracting *useful information* from it has shown to be usually a nontrivial task.

Complex systems don't allow for detailed theories to emerge *before* observation; consequently, they don't allow for problem-specific shaping of the experimental designs beyond some relatively trivial constraints. The result is that we often find ourselves with enormous amounts of data that could very plausibly contain enough information to characterize in full detail the functioning of a system, but our understanding (whatever we precisely want to mean by this) of such system does not progress much, due to our fundamental incapability to spot what is *relevant* inside this data.

This being said, the aforementioned absence of prior models often puts us in the situation where the most meaningful thing to do is to try and characterize the statistical dependencies between observed quantities: in other words, we aim for reconstruction of the *generative probability distribution* from which supposedly our sample has been "drawn".

Implicit in this line of thought is that we are, in general, looking at a set of sample points that we suppose to be *independent and identically distributed (i.i.d).* [1] These being the premises, let's set up things.

[1] Infinite digressions could be possible here, from the most abstract ones about the many pitfalls of "*iid* reasoning" when dealing with real phenomena (cfr. the beautiful discussion in chapter 3 of [7]) to many more problem-specific ones; we will just ignore all of these: the hope is that we can resort to Wigner's *"skill and ingenuity"* at least for

1.1 Statistical model selection

Suppose we have collected data: $\widehat{x}_N = (x^{(1)}, \cdots, x^{(N)})$, $x^{(i)} \in \Omega \ \forall i \in [1, N]$ consisting of N observations $x^{(i)}$ that we suppose i.i.d $\sim \rho^* : \Omega \to [0, 1]$. The process by which one obtains an estimate $\hat{\rho}$ for the function ρ^* (the *generative distribution*) starting from data is usually divided in two steps:

- *Choice of a parametric family* (model) for our putative $\hat{\rho}$ to live in;
- *Parameter estimation*, to be performed once we've taken for granted a parametric family to work with.

Let's elucidate what we mean with a classic example: suppose we are given a set of points on the xy plane like the one in figure 1.1:

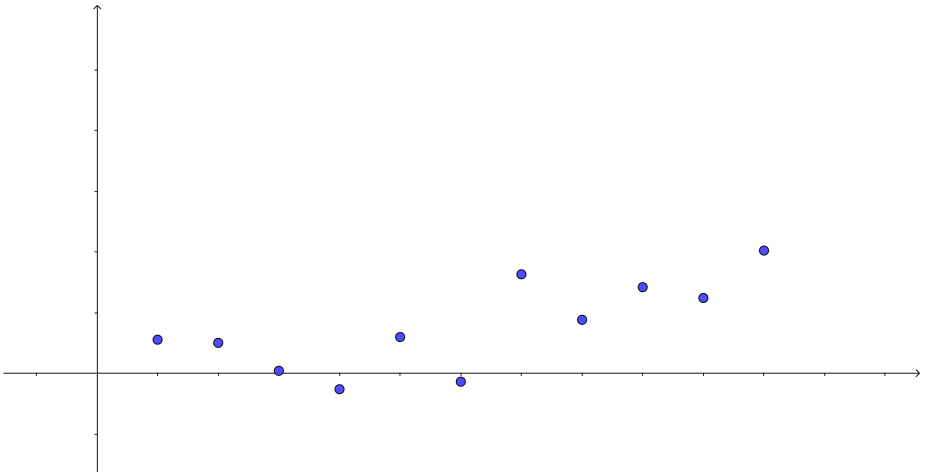

Figure 1.1: *A sample of observations on the xy plane*

What we will do in a case like this is tipically opt for a noisy linear model.

$$y_i = \alpha + \beta x_i + \eta_i \tag{1.1}$$

with $\{\eta_i\}$ modeled to be iid and Gaussian, so that:

$$\rho(y|x; \alpha, \beta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\alpha-\beta x)^2}{2\sigma^2}} \tag{1.2}$$

The classic recipe for the estimation of α and β prescribes then that we look for the values $\hat{\alpha}$,$\hat{\beta}$ maximizing the *likelihood*:

$$P(\hat{x}_N|\alpha, \beta, \sigma) = \prod_{i=1}^{N} \rho(y_i|x_i; \alpha, \beta, \sigma) \tag{1.3}$$

what's enough to be reasonably confident that no highly relevant, data-point dependent, latent variables have been left outside of our analyses.

Showing that $\hat{\alpha}$,$\hat{\beta}$ do not depend on σ and finding their values is a classic exercise in basic statistical courses. What is seldom drawn attention to is the model selection stage, that we implicitly performed the moment we decided to resort to a linear function. We could have actually chosen an higher degree polynomial: the increase in the number of parameters of our model allows for its higher espressivity, resulting in higher scaled loglikelihoods and smaller mean square errors; yet, it is intuitively clear that this gain in interpolation power (the so-called *goodness of fit*) masks a huge loss in *generalization* capability: if we found our high-degree "best fit" parameters for N data points, and then drawed another independent observation, chances are that the added data point's position would be predicted very poorly, surely worse than under a linear assumption; the reason, in this simple case, is evident: beyond first order, we are *fitting the noise*.

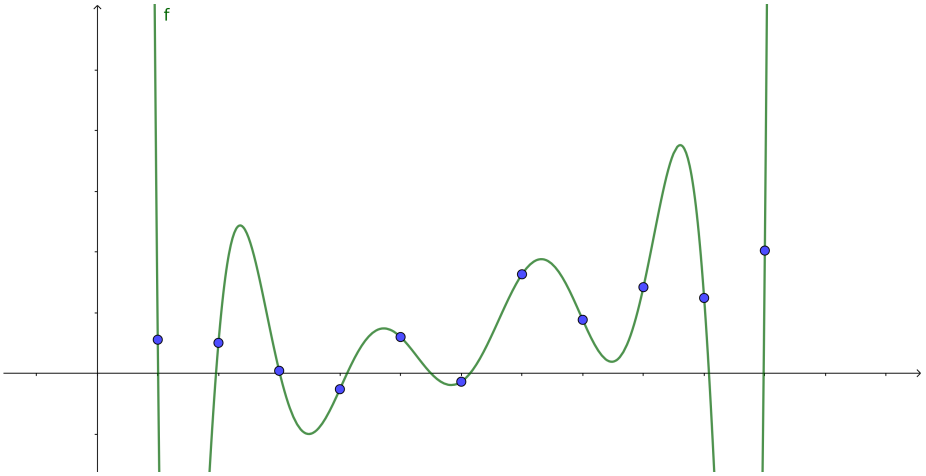

Figure 1.2: *"Perfect fit" using a high-dimensional model*

Let's take a moment to ponder on the fact that, weren't we in the lucky case of simultaneously being in two dimensions (so that things are easily visualizable) and having an "evidently" linear data (so that things are easily interpretable) we would not have had any predefined strategy on how to choose the correct model.
From now on, in the present chapter, our task will thus be that of discussing model selection, intended as a family of tools and strategies aimed at being able to find, in a nonparametric framework, models which have a nice goodness of fit while securing ourselves against their potential *overfitting* ([8]).
From now on, we will work in a more formal fashion, and assume the reader is familiar with basic instruments in probability theory, especially in its Bayesian framework (as a reference, see for instance [9] [10]).

1.2 Hypotheses testing and distinguishability

We will closely follow the exposition in [11] and [12] in this section. We will also use classic results at the interface between statistics and information theory, a detailed exposition of which can be found for instance in [13].

1.2.1 Distinguishing probability distributions

Consider the space $\mathbb{P}$ of all probability distributions on a discrete alphabet S. Take two points P_1, P_2 of this space, i.e. two probability distributions on S, and regard them as alternative statistical hypotheses for an unknown distribution P^* from which we have means of drawing data samples $\hat{s}_N$. Once we have fixed the number N of sample points to be drawn, it is a meaningful question to ask ourselves wether *we can effectively distinguish between P_1 and P_2, as candidates for the generative distribution of the data, by observing a number N of data points.* In other words, the main question is: are N i.i.d. sample points enough to determine the correct hypothesis with a "good" (arbitrarily set) confidence? In order to assess this question, let's first recall the usual definition of *error probabilities* in hypothesis testing:

$$\alpha = P_1(\overline{A}) \qquad \beta = P_2(A) \tag{1.4}$$

Here, $A \in S^N$ is the *acceptance region* of P_1, i.e. the subset of samples for which we accept P_1, and $\overline{A} = \mathbb{P} \setminus A$ is its *rejection region*; α then represents the probability of erroneously rejecting P_1, and β the probability of erroneously accepting it. Ideally, we would like to make both these quantities as small as possible. The quantitative definition of a possible acceptance region is naturally given by setting a threshold T on the posterior ratio:

$$A_T = \left\{ \hat{s}_N \mid \frac{P_1(\hat{s}_N)}{P_2(\hat{s}_N)} > T \right\} \tag{1.5}$$

Here, once specified both a threshold T, we can absorb the information of the latter in the former, getting a "likelihood ratio"-like criterion with a modified threshold:

$$A_T = \left\{ \hat{s}_N \mid \frac{P_1(\hat{s}_N)}{P_2(\hat{s}_N)} > \widetilde{T}_\pi \right\}, \qquad \widetilde{T}_\pi = \frac{T(1-\pi)}{\pi} \tag{1.6}$$

Another useful representation of this set can be obtained by first taking logarithms:

$$A_T = \left\{ \hat{s}_N \mid \log\left(\frac{P_1(\hat{s}_N)}{P_{\hat{s}_N}(\hat{s}_N)} \right) > \log\left(\frac{P_2(\hat{s}_N)}{P_{\hat{s}_N}(\hat{s}_N)} \right) + \log(\widetilde{T}_\pi) \right\} \tag{1.7}$$

and then averaging over $P_{\hat{s}_N}$, it being the *empirical distribution* (also known as *type*) of the sample:

$$A_T = \left\{ \hat{s}_N \mid D_{KL}(P_{\hat{s}_N}||P_1) < D_{KL}(P_{\hat{s}_N}||P_2) + \log \widetilde{T}_\pi \right\} \tag{1.8}$$

From this last expression we get a natural geometrical interpretation of the acceptance criterion as an evaluation of which hypothesis is "closest" (in the D_{KL} sense) to the empirical distribution.
If we now look for the empirical distribution $P_{\underline{X}}$ which, while belonging to A_T, minimizes $D_{KL}(P_{\underline{X}}||P_2)$, we find [13]:

$$P^{\lambda}_{\underline{X}}(s) = \frac{P_1^{\lambda}(s)P_2^{1-\lambda}(s)}{\sum_r P_1^{\lambda}(r)P_2^{1-\lambda}(r)} \tag{1.9}$$

where the Lagrange multiplier λ has to be assigned a value so as to match the required threshold T.
Equation (1.9) can be interpreted as a "straight line" linking P_1 and P_2; a basic result of large deviation theory, *Sanov's theorem* (see again [13]), tells us that if we fix $\alpha = \epsilon$ then our lowest possible value for β is

$$\beta_N = 2^{-ND_{KL}(P_\lambda||P_2)+o(\epsilon)} \tag{1.10}$$

1.2.2 Models

Let's get back to $\mathbb{P}$. The choice of a *parametric family*

$$p(\cdot|\vec{\theta}) = f(\cdot, \vec{\theta}) : S \longrightarrow \mathbb{R}$$

of normalized functions identifies, by letting θ vary continuously inside some parametric domain $\Theta \subset \mathbb{R}^n$, a subset $\mathfrak{M} \subset \mathbb{P}$ that we'll call a *parametric model*. Let's now take two close points

$$\vec{\theta_0}, \vec{\theta} \qquad |d\vec{\theta}| = |\vec{\theta} - \vec{\theta}_0| \ll 1$$

on this model, and ask ourselves again wether we are able to distinguish between them. In particular, we are interested in the case in which these two distributions are the maximum likelihood distributions relative to two "similar" samples.
Fixed a confidence level $\epsilon \ll 1$, it is not guaranteed that $\beta_N \ll \epsilon$, since $D_{KL}(P_1||P_2) \ll 1$ now as well.
There will be then, in general, a region $V_{\epsilon,N}(\vec{\theta_0}) \subset \mathfrak{M}$, surrounding $\vec{\theta}_0$ composed of distributions that cannot be confidently distinguished on the basis of N sample points only. To get some sort of measure of the size of this region, we can use Stein's lemma:

$$\epsilon > 2^{-ND_{KL}(P_{\vec{\theta}_0}||P_{\vec{\theta}})} \tag{1.11}$$

Expanding the Kullback-Leibler divergence around $\vec{\theta_0}$:

$$\begin{aligned}
D_{KL}(P_{\vec{\theta_0}}||P_{\vec{\theta}}) =& -\sum_s \left(P_{\vec{\theta_0}}(s) \log \frac{P_{\vec{\theta_0}+\vec{d\theta}}(s)}{P_{\vec{\theta_0}}(s)} \right) \\
=& -H[P_{\vec{\theta_0}}] - \sum_s \left(P_{\vec{\theta_0}}(s) \log P_{\vec{\theta_0}+\vec{d\theta}}(s) \right) \\
\approx& -H[P_{\vec{\theta_0}}] + H[P_{\vec{\theta_0}}] - \sum_s \left(P_{\vec{\theta_0}}(s) \sum_i d\theta_i \frac{\partial}{\partial \theta_i} \left[\log P_{\vec{\theta}} \right]_{\theta_0} \right) + \\
& - \sum_s \left(P_{\vec{\theta_0}}(s) \sum_{i,j} \frac{d\theta_i d\theta_j}{2} \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \left[\log P_{\vec{\theta}} \right]_{\theta_0} \right) \\
=& \sum_i d\theta_i \sum_s P_{\vec{\theta_0}}(s) \frac{1}{P_{\vec{\theta_0}}(s)} \frac{\partial}{\partial \theta_i} \left[P_{\vec{\theta}} \right]_{\theta_0} + \\
& - \sum_{i,j} \frac{d\theta_i d\theta_i}{2} \sum_s P_{\vec{\theta_0}}(s) \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \left[\log P_{\vec{\theta}} \right]_{\theta_0} \\
=& \sum_i d\theta_i \frac{\partial}{\partial \theta_i} \left[\mathbb{E}_s(1) \right] - \frac{1}{2} \sum_{i,j} d\theta_i d\theta_j \, \mathbb{E}_s[\partial_i \partial_j \log P_{\vec{\theta_0}}] \\
=& \frac{1}{2} \sum_{i,j} d\theta_i d\theta_j \mathbb{J}_{ij}(\vec{\theta_0})
\end{aligned} \tag{1.12}$$

where we identified the *Fisher information*:

$$\mathbb{J}_{ij}(\vec{\theta}) = -\mathbb{E}_s[\partial_i \partial_j \log P_{\vec{\theta}}] \tag{1.13}$$

We finally get:

$$\epsilon > e^{-\frac{N}{2} \vec{\delta\theta}^T \mathbb{J}(\vec{\theta_0}) \vec{\delta\theta}} \tag{1.14}$$

leading to:

$$\vec{\delta\theta}^T \mathbb{J}(\vec{\theta_0}) \vec{\delta\theta} < -\frac{2}{N} \log \epsilon \tag{1.15}$$

Geometric interpretation

This inequality is, geometrically, the definition of a small ellipsoid's interior. Its interpretation must be that of a tiny set of probability distributions between which we cannot effectively distinguish, at chosen sample size N and threshold T. In this respect, the Fisher information matrix turns out to be closely related to the notion of what *resolution* we can afford in the space of putative probability distributions.

This result will be crucial for us, since we shortly derive from it a sound solution to the old problem of what prior probabilities $P_0(\vec{\theta_{\mathcal{M}}}|\mathcal{M})$ should be chosen for inference when working with continuously-parametrized families of distributions.

1.2.3 Counting probability distributions

The ellipsoid we just managed to define has volume:

$$V_{\epsilon,N}(\vec{\theta_0}) \propto \frac{1}{\sqrt{\det J(\vec{\theta_0})}} \tag{1.16}$$

so that the number of such ellipsoids in a small cube $\vec{\Delta}\theta_0$ surrounding $\vec{\theta_0}$ will be:

$$\Delta\mu(\vec{\theta_0}) = \frac{\vec{\Delta}\theta}{V_{\epsilon,N}(\vec{\theta_0})} \propto \sqrt{\det J(\vec{\theta_0})} \cdot \vec{\Delta}\theta \tag{1.17}$$

In the limit $N \to \infty$ these ellipsoids become infinite and "dense", but the ratio between the number of them in a small region and the number of them in the whole parametric family converges ([11]), defining a "distinguishability measure" over $\mathfrak{M}$:

$$d\mu(\vec{\theta}) = \left[\frac{\sqrt{\det J(\vec{\theta})}}{\int_{\mathfrak{M}} \sqrt{\det J(\vec{\theta'})} d\vec{\theta'}}\right] d\vec{\theta} \tag{1.18}$$

We thus get an insightful result:
Given a sample size N, the space of probability distributions is effectively partitioned in "distinguishable" cells.
In the limit $N \to \infty$ these define a natural measure on $\mathfrak{M}$ weighting equally parametric regions containing the same number of distinguishable distributions. [2]
Thus, a distribution over $\vec{\theta}$ which is uniform with respect to this measure is effectively a uniform distribution over all distinguishable distributions. Such distribution is known as "Jeffreys prior":

$$\rho_0(\vec{\theta})d\vec{\theta} = \frac{\sqrt{\det J(\vec{\theta})}}{\int_{\mathfrak{M}} \sqrt{\det J(\vec{\theta'})} d\vec{\theta'}} d\vec{\theta} \tag{1.19}$$

It acts as an "ignorance" prior in the sense just described, and enjoys many useful properties - above all, full invariance under reparametrization. We'll now see how this object and his properties affect model selection.

[2] For a full differential-geometrical treatment of models as Riemannian manifolds under a Fisher Information defined metric, see [14]

1.3 Bayesian Model Selection

One of the most popular, and theoretically sound, recipes for performing model selection is that of simply applying Bayes rule. If we define a set of possible models to choose between, we can then compute the conditional probabilities of different models, conditioned on the observed sample:

$$P(\mathcal{M}|\hat{x}_N) = \frac{P(\hat{x}_N|\mathcal{M})P(\mathcal{M})}{P(\hat{x}_N)} \tag{1.20}$$

We're not making any kind of assumptions on the nature of models yet: if we choose to regard models as particular parametric families of distributions (e.g. the gaussian family, the mixture family, the "finite polynomials of order k" family...), then we can expand at the right hand side:

$$P(\mathcal{M}|\hat{x}_N) = \frac{P(\mathcal{M})}{P(\hat{x}_N)} \int_{\mathcal{M}} d\theta_{\mathcal{M}} P(\hat{x}_N|\theta_{\mathcal{M}}, \mathcal{M})\rho(\theta_{\mathcal{M}}|\mathcal{M}) \tag{1.21}$$

If the data points are supposed to be iid, the likelihood at the integrand factors, giving:

$$\begin{aligned} P(\hat{x}_N|\theta_{\mathcal{M}}, \mathcal{M}) =& \prod_{i=1}^{N} P(\hat{x}^{(i)}|\theta_{\mathcal{M}}, \mathcal{M}) = e^{\sum_{i=1}^{N} \log P(\hat{x}^{(i)}|\theta_{\mathcal{M}}, \mathcal{M})} \\ =& e^{N\overline{\log P(\hat{x}|\theta_{\mathcal{M}}, \mathcal{M})}} \equiv e^{N\overline{\mathcal{L}_{\hat{x}}(\theta)}} \end{aligned} \tag{1.22}$$

in which the horizontal line denotes an arithmetic mean among the N observations. When N is large, we can expand $\overline{\mathcal{L}_{\hat{x}}(\theta)}$ to second order around its maximum, to later use saddle point integration:

$$P(\mathcal{M}|\hat{x}_N) \approx \frac{P(\mathcal{M})}{P(\hat{x}_N)} e^{N\overline{\mathcal{L}_{\hat{x}}(\theta^*)}} \int_{\mathcal{M}} d\theta_{\mathcal{M}} \rho(\theta_{\mathcal{M}}|\mathcal{M}) e^{-\frac{N}{2}\sum_{a,b}^{|\mathcal{M}|} d\theta_a d\theta_b (-\partial_a \partial_b \overline{\mathcal{L}_{\hat{x}}(\theta)})} \tag{1.23}$$

Let's now drop $P(\hat{x}_N)$ since it serves merely as a normalization factor here, and assume we have no information that allows to prefer a model over another, so that our prior $P(\mathcal{M})$ will be chosen as uniform (and thus dropped out as well); the problem would now remain of choosing a suitable prior $P(\theta_{\mathcal{M}}|\mathcal{M})$ over the parameter space defined by model $\mathcal{M}$...if we hadn't just solved this problem via the introduction of our distinguishability metric!

Plugging 1.1.19 in our present expression yields then:

$$
\begin{aligned}
P(\mathcal{M}|\hat{x}_N) \propto & \frac{e^{N\overline{\mathcal{L}_{\hat{x}}(\theta^*)}}}{\int_{\mathcal{M}} \sqrt{\det \mathbb{J}(\vec{\theta'})} d\vec{\theta'}} \int_{\mathcal{M}} d\theta_{\mathcal{M}} \sqrt{\det \mathbb{J}(\vec{\theta})} e^{-\frac{N}{2} \sum_{a,b}^{|\mathcal{M}|} d\theta_a d\theta_b (-\partial_a \partial_b \overline{\mathcal{L}_{\hat{x}}(\theta)})} \\
= & \frac{\sqrt{\det \mathbb{J}(\theta^*)}}{\sqrt{\det J_{\hat{x}_N}(\theta^*)}} \frac{e^{N\overline{\mathcal{L}_{\hat{x}}(\theta^*)}}}{\int_{\mathcal{M}} d\theta_{\mathcal{M}} \sqrt{\det \mathbb{J}(\theta_{\mathcal{M}})}} \left(\frac{2\pi}{N}\right)^{\frac{|\mathcal{M}|}{2}} \\
= & \exp\left[N\overline{\mathcal{L}_{\hat{x}}(\theta^*)} - \frac{|\mathcal{M}|}{2} \log\left(\frac{N}{2\pi}\right) - c_{\mathcal{M}}^{BMS} - r_{\mathcal{M}}^{BMS}(\hat{x}_N)\right]
\end{aligned}
\tag{1.24}
$$

where we defined:

$$
c_{\mathcal{M}}^{BMS} = \int_{\mathcal{M}} d\theta_{\mathcal{M}} \sqrt{\det \mathbb{J}(\theta_{\mathcal{M}})} \tag{1.25}
$$

to be the *geometric complexity* of the model, and

$$
r_{\mathcal{M}}^{BMS}(\hat{x}_N) = \log \sqrt{\frac{\det J_{\hat{x}_N}(\theta^*)}{\det \mathbb{J}(\theta^*)}} \tag{1.26}
$$

to be its *relative complexity* with respect to the drawn sample.
There are a few things to notice here. The four terms in the last exponent have different behavior for growing N:

- For N very large, as expected, our posterior estimate for the probability of a given model is completely driven by maximization of the scaled loglikelihood $\overline{\mathcal{L}_{\hat{x}}} = \frac{1}{N} \sum_{i=1}^{N} \log P(\hat{x}^{(i)}|\theta_{\mathcal{M}}, \mathcal{M})$;

- as N shrinks, the second term, with its logarithmic dependence upon N, becomes more and more relevant: this term, known in the literature as Bayesian Information Criterion (BIC) and widely used in practice when addressing general model selection problems, acts as a penalization directly proportional to the number of parameters in the chosen parametric family M. It has the net effect of shifting our estimate away from models which overfit the data due to excessive degrees of freedom (remember our first example!), and guiding us towards "simpler" models (by a first naive notion of simplicity as "number of free parameters").

- As N shrinks even more towards the undersampling regime, the terms $c_{\mathcal{M}}^{BMS}$ and $r_{\mathcal{M}}^{BMS}(\hat{x}_N)$ come into play.

1.4 Geometric interpretation of $c_{\mathcal{M}}^{BMS}$ and $r_{\mathcal{M}}^{BMS}(\hat{x}_N)$

Let's start from the geometric complexity $c_{\mathcal{M}}^{BMS} = \int_{\mathcal{M}} d\theta_{\mathcal{M}} \sqrt{\det \mathbb{J}(\theta_{\mathcal{M}})}$. As for what we've seen in section 1.2, the geometric interpretation of this term is evident: its value is *directly proportional to the number of distinguishable probability distributions contained in the parametric family* $\mathcal{M}$. In this sense we will say that a model is geometrically more complex if it is capable of reproducing a large number of distributions; this must be thought of as an effect *beyond* the one of mere dimensionality, since the complexity relative to the model dimension has already been taken care of in the BIC term.

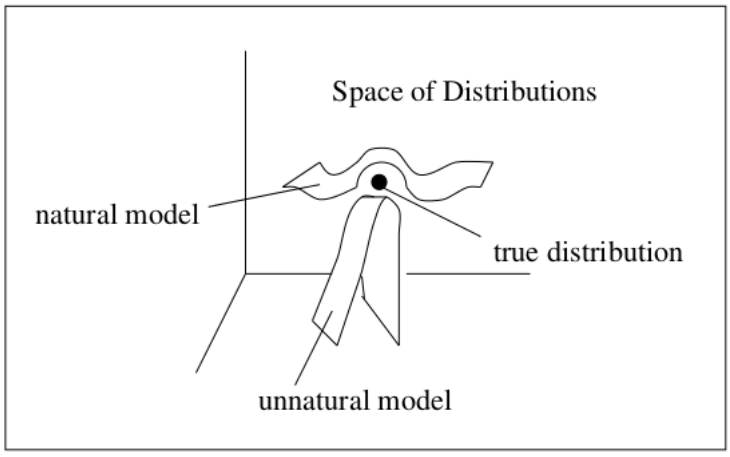

Figure 1.3: *"Naturalness" of models.*
[*Image from: V. Balasubramanian, "A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions"* ([[11]])]

The relative complexity $r_{\mathcal{M}}^{BMS}(\hat{x}_N)$ is more of a *local* quantity: the way it emerges during the evaluation of the above integral suggests that its geometric meaning is that of a ratio between the number of distinguishable distributions on the model manifold in the close surroundings of the maximum likelihood point and the total numer of the model's distinguishable distributions. In other words , it measures *what fraction of the distributions seen by the model lies "close" to the best one.* In [11], the author refers to this property of "accumulation" around the true estimate as *naturalness* of a model. Again: we could say that this term measures the robustness of the model, intuitively defined as its sensitivity to the precise choice of parameters. Consider again our first example: in that case the high-degree model was not at all *robust*: in that case adding an additional data points is of maximum likelihood parameters *a lot.* This great sensitivity of the parameters with respect to additional data implies that the model is very sensitive

to sampling noise, and thus will exhibit potentially very different estimates for different realizations of a sample draw. This behavior is the same described in recent work on so-called *sloppy models* (see for instance [15], [16]), and is peculiar of models that overfit the data. Notice that if the true distribution lies within the considered model family, $J_{\hat{x}_N}(\theta^*)$ approaches $\mathbb{J}(\theta^*)$ as N grows large, thus making this term negligible; if, conversely, the true distribution lies outside of the model chosen, this term acts as a precious proxy for robustness of our approximation of this true distribution.

1.5 Recap

Let's repeat, in summary, what we have achieved so far:

- We have raised the issue of model selection, and understood its importance, especially in application to complex systems, as the stage of inference when we should secure ourselves as much as we can from the risk of *overfitting*;
- We have introduced basic notions and results about distinguishability of hypotheses given limited datasets; an information-theoretic approach has permitted us to obtain a well defined, quantitative *distinguishability metric* which not only allows us to solve the problem of choosing priors on parameter spaces, but also enables many geometrical considerations and interpretations about subsequent results and objects;
- We have followed a plain Bayesian procedure to try and quantify the asymptotic behavior of the posterior distribution of models; this way we have retrieved the classic BIC criterion, *plus* two more *complexity penalties* arising from the geometric features of the models chosen.

In the following chapter, we will employ the concepts just described in order to justify the choice of a particular class of systems to perform model selection in, namely *spin systems with interactions of arbitrary order*, for the general statistical inverse problem in complex scenarios.

Chapter 2

Spin models & their complexity

Introduction: modelling with spin systems

Nowadays, we have access to direct observation of microscopic degrees of freedom of complex systems in a variety of cases. The most notable examples are found in biology, where for instance today we have:

- the possibility of simultaneous recording of multiple neuron's activity *in vivo* via multi-electrode arrays;
- the rapidly advancing DNA and mRNA sequencing technologies;
- the possibility to determine 3D structure of proteins via X-ray cristallography;

To treat these phenomena with the machinery of statistical physics means to enforce statistical mechanical models, then estimate models' parameters from data.

A comprehensive review of statistical inverse problems arisen in the Big Data era can be found in [17]. A paradigmatic case is that of models constituted of asseblies of *binary* interacting units. This is a natural setting when, for instance, we work with neurons' activity patterns (where the two states are "firing" and "not firing" for each neuron); the astonishing spread of artificial neural based architectures in the machine learning practice has made binary spin models a crucial framework to be investigated and understood in detail.
What is usually found in the statistical learning literature and practice is that whenever we adopt a binary spin model to describe some system, we restrict ourselves to at most *pairwise* interactions [1]; this means that the

[1] yet, see [18], [19]

Hamiltonian for a system of n spins will be of the form

$$\mathcal{H}(\mathbf{s}) = \sum_{i=1}^{n} h_i s_i + \sum_{i=1}^{n} \sum_{j<i} J_{ij} s_i s_j$$

and the corresponding Boltzmann distribution will look like:

$$\rho(\mathbf{s}) = \frac{1}{\mathcal{Z}_{\vec{h},\vec{J}}} e^{\beta(\sum_i h_i s_i + \sum_{i \neq j} J_{ij} s_i s_j)} \tag{2.1}$$

We recall that for distributions in the exponential family: $P(s) = e^{\sum_j c_j f_j(s)}$ the classic result is that the functions $f(s)$ are the *sufficient statistics* of the model; this means that their average over the sample is a quantity containing all the possible information about the associated parameters c_j - and through them, about the generative distribution (see, for instance, [20]). In this perspective, the choice of including only one and two body interactions in 2.1 entails a strong prejudice about the fact that only one and two point correlations are relevant for characterization of the system's behavior. This sharp prior confidence in which are the relevant informations contained in the data and which are not is often not justified by the preliminar information we have at our disposal. The reasons why at-most-pairwise models are so popular are simple:

- there is a *lot* of theoretical work that can be borrowed from orthodox statistical mechanical theory for pairwise-interacting systems;
- the number of parameters to be inferred (i.e. the *dimensionality* of the model) grows very fast as we raise the maximum order of interactions allowed. So if we have n degrees of freedom for a non interacting n-spin model, we'll have $n + \frac{n(n-1)}{2} \sim o(n^2)$ for a single+pairwise model, $o(n^3)$ for an at-most-triplewise one, and so on. It is usually enough of a struggle to perform stable inference in the pairwise case, so that a complete treatment looks "nice, but not affordable".
- pairwise models often *work very well*; a detailed discussion of this phenomenon of "pairwise sufficiency" be found in [21] (also note that, as argued in [22], adding pairwise-correlated latent units allows to mimic interactions of every order).
- the great success of machine learning architectures composed on large assemplies of pairwise interacting neurons has stimulated a lot of research work in this particular setting.

Still, from an abstract point of view, there's no justification in this choice.

In this chapter we will argue that the attempt to give formal justification to the choice of pairwise models on the ground of Bayesian Model Selection

protocols is doomed to fail. We will do this by exhamining the structure of the "any-order-interactions" general model while enforcing complete prior uninformativity.
We will follow mainly [4].

2.1 Generalities

2.1.1 Spin systems with interactions of arbitrary order

A *binary spin system* of size $n \in \mathbb{N}$ is an assembly of n random binary variables:

$$(s_1, \cdots, s_n) = \mathbf{s}$$
$$s_j = \pm 1 \quad \forall j \in \{1, \cdots, n\}$$
$$\mathbf{s} \in \{-1, +1\}^n \equiv \mathcal{S}_n$$

Definition

If we want to take into account interactions of arbitrary order between spins, and to do so without structurally enforcing any kind of simmetry in our system, we can resort to this basis of *generalized spin operators*:

$$\phi^{\mu}(s) = \prod_{i \in \mu} s_i \tag{2.2}$$

Where $\mu \in \mathcal{P}(\{1, 2, \cdots, n\})$ identifies a subset of spins (e.g. $\phi^{(1,2,7)}(s) = s_1 s_2 s_7$).
Our generic spin Hamiltonian will then be a linear combination

$$\mathcal{H}(s) = \sum_{\mu} g^{\mu} \phi^{\mu}(s)$$

of these operators; we can retrieve from it the usual cases (e.g. Ising, Curie-Weiss, REM...) simply by enforcing constraints to be obeyed by the parameters g (e.g. requiring $g^{\mu} = 0$ for any μ comprising more than two spins gives a pairwise model).

Indexing

To avoid awkward notations for the indices μ, we represent each subset of $\{1, \cdots, n\}$ as a binary string, with "1"s placed in positions denoting (counting from right to left) spins belonging to the subset chosen; we then assign to μ the natural value whose binary representation coincides with that string.
For instance, if n=9:

$$\phi^{(1,2,7)}(\mathbf{s}) \to (1, 2, 7) \to 001000011 \to 2^6 + 2^1 + 2^0 = 67 \to \phi^{67}(\mathbf{s})$$

$$\phi^{(1,2,7)}(\mathbf{s}) \equiv \phi^{67}(\mathbf{s})$$

so that $\mu = 67$ will denote the interaction between the binary units $(1, 2, 7)$. This means that now we can just use natural numbers: $\mu \in \{0, 1, \cdots, 2^n - 1\}$; here we also included $\mu = 0$, corresponding to the *identity operator*:

$$\phi^0(\mathbf{s}) = 1 \quad \forall \mathbf{s} \in \mathcal{S}$$

Let's call Ω_n the set of operators just defined

Properties

We'll now highlight some interesting properties of the spin operator family. First of all, it is trivial to verify that Ω_n is closed under multiplication:

$$\phi^\mu(\mathbf{s})\phi^\nu(\mathbf{s}) = \prod_{i\in\mu}\prod_{j\in\nu} s_i s_j = \prod_{\substack{k\in\mu\\k\in\nu}} s_k^2 \prod_{\substack{i\in\mu\ j\in\nu\\i\notin\nu\ j\notin\mu}} s_i s_j = \prod_{\substack{i\in\mu\ j\in\nu\\i\notin\nu\ j\notin\mu}} s_i s_j = \phi^{\mu\oplus\nu}(\mathbf{s}) \tag{2.3}$$

where we denote by $\oplus$ the XOR operator between the binary strings representing the two operators.
Also:

$$\phi^\mu(\mathbf{s})\phi^0(\mathbf{s}) = \phi^\mu(\mathbf{s}) \quad \forall \mu \in \Omega_n \tag{2.4}$$

$$\phi^\mu(\mathbf{s})\phi^\mu(\mathbf{s}) = \phi^{\mu\oplus\mu}(\mathbf{s}) = \phi^0(\mathbf{s}) = 1 \tag{2.5}$$

from which is evident that Ω_n constitutes an Abelian group under multiplication of operators; since every element in it is its own inverse, this group is isomorphic to ${\mathbb{Z}_2}^n$.
The two most useful properties of these operators are their *orthogonality*:

$$\sum_{\mathbf{s}\in\mathcal{S}} \phi^\mu(\mathbf{s})\phi^\nu(\mathbf{s}) = \sum_{\mathbf{s}\in\mathcal{S}} \phi^{\mu\oplus\nu}(\mathbf{s}) = 2^n \delta_{\mu\oplus\nu,0} = 2^n \delta_{\mu,\nu} \tag{2.6}$$

and *completeness*:

$$\begin{aligned}\sum_\mu \phi^\mu(\mathbf{s})\phi^\mu(\mathbf{r}) &= \sum_\mu (\prod_{i\in\mu} s_i)(\prod_{j\in\mu} r_j) = \sum_\mu \prod_{i\in\mu} s_i r_i \\ &= \sum_\mu \phi^\mu(s_1 r_1, s_2 r_2, \cdots, s_n r_n) \\ &= 2^n \prod_i^n \delta_{s_i r_i, 1} = 2^n \delta_{\mathbf{s},\mathbf{r}}\end{aligned} \tag{2.7}$$

This last relation in particular is crucial, since it shows us how we can use the $\{\phi^\mu\}$ basis to represent any desired function $f(\mathbf{s})$ of the spin configurations:

$$\begin{aligned} f(\mathbf{s}) &= \sum_{\mathbf{r}} \delta_{\mathbf{s},\mathbf{r}} f(\mathbf{r}) = \frac{1}{2^n} \sum_{\mathbf{r}} \sum_{\mu} \phi^\mu(\mathbf{s}) \phi^\mu(\mathbf{r}) f(\mathbf{r}) \\ &= \sum_{\mu} \phi^\mu(\mathbf{s}) \sum_{\mathbf{r}} \frac{\phi^\mu(\mathbf{r}) f(\mathbf{r})}{2^n} = \sum_{\mu} g^\mu \phi^\mu(\mathbf{s}) \end{aligned} \tag{2.8}$$

It is also useful to take note of the following two relations, arising from the orthogonality and completeness properties

$$\sum_{\mathbf{s}\in\mathcal{S}} \phi^\mu(\mathbf{s}) = \sum_{\mathbf{s}\in\mathcal{S}} \phi^\mu(\mathbf{s}) \phi^0(\mathbf{s}) = 2^n \delta_{\mu,0} \quad \forall \phi^\mu \in \Omega_n \tag{2.9}$$

$$\sum_{\mu} \phi^\mu(\mathbf{s}) = \sum_{\mu} \phi^\mu(\mathbf{s}) \phi^\mu(+,\cdots,+) = 2^n \delta_{\mathbf{s},(+,\cdots,+)} = 2^n \prod_{i}^{n} \delta_{s_i,1} \quad \forall \mathbf{s} \in \mathcal{S} \tag{2.10}$$

the second one being true due to the fact that $\phi^\mu(+,\cdots,+) = 1 \ \forall \mu \in \Omega_n$.

Generating sets, independent sets

We've seen above that Ω_n, as a group equipped with the XOR composition operator, is isomorphic to $\mathbb{Z}_2^n$. A little pondering on this induces us to introduce the notion of a *generating set* of operators.
We will call a collection of operators "generating for Ω_n" if every element of Ω_n is obtainable by an arbitrary number of XOR compositions of elements of this collection.
The simplest generating set thinkable is the set of all "monomial" operators:

$$\mathcal{I}_1 = \{\phi^{10\cdots0}, \phi^{01\cdots0}, \cdots, \phi^{00\cdots1}\} \supset \Omega_n$$

from which it is trivial to obtain any other one.

We'll call a collection of operators *independent* if no operator belonging to it can be obtained by XOR compositions of the others.
It is clear that our "monomial" example set $\mathcal{I}_1$ is independent. It is also clear that there is no other element that can be added to this collection without compromising its independency.
In general, a set of n independent operators is automatically a generating set of Ω_n.

In the following, it can happen that we do not refer to the entire space Ω_n, but to a subset $\mathcal{M}$ (a model) of it. The definition of "generating set of $\mathcal{M}$ will then be the one naturally induced from the general case.

2.1.2 Spin models

A particular choice of a family of operators $\mathcal{M} \subseteq \Omega_n$ identifies a parametric family for the system's Hamiltonian to live in, and thus, specifies a *model*. Given model $\mathcal{M}$, we construct the Boltzmann / maximal entropy probability distribution for our spin variables:

$$P(\mathbf{s}|\mathbf{g}, \mathcal{M}) = \frac{1}{\mathcal{Z}_{\mathcal{M}}(\mathbf{g})} e^{\sum_{\mu \in \mathcal{M}} g^\mu \phi^\mu(\mathbf{s})} \tag{2.11}$$

To lighten the notation, we'll define:

$$g^0(\mathbf{g}) = -\log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}) \tag{2.12}$$

so that:

$$P(\mathbf{s}|\mathbf{g}, \mathcal{M}) = e^{g^0 + \sum_{\mu \in \mathcal{M}} g^\mu \phi^\mu(\mathbf{s})} = e^{\sum_{\mu \in \widetilde{\mathcal{M}}} g^\mu \phi^\mu(\mathbf{s})} \tag{2.13}$$

where $\widetilde{\mathcal{M}} \equiv \mathcal{M} \cup \{\phi^0\}$.
From now on, we will use the symbol $\mathcal{M}$ in both cases: it will be clear from context if we are implicity or explicitly expressing the partition function.

A model will be called *nondegenerate* if it contains no additional constraints to be fulfilled by its parameters (any fully disordered system is nondegernerate in this sense); *degenerate* models are those in which we enforce some such additional constraints (e.g. the usual Ising model is degenerate, since we require that all pairwise interactions have the same magnitude:

$$g_{12} = g_{13} = \cdots = g_{23} = \cdots = J$$

2.2 Bayesian Model Selection on spin models

It is easy to see that the number of different nondegenerate models possible is $2^{|\Omega_n|} = 2^{2^n - 1}$, which is huge; the core question therefore is: **to what extent we can, if given a sample $\hat{s}_N$ of N observed (and supposed iid $\sim \rho_{gen}$) spin configurations, significantly select a model for the distribution ρ_{gen} that generated the sample?**

A Bayesian estimate

Let's again look for the posterior distribution af models conditioned on the observed data and repeat the computation we already went through in Chap-

ter 1, this time with details specific of this particular case:

$$
\begin{aligned}
P(\mathcal{M}|\hat{\mathbf{s}}_N) &= \frac{P(\hat{\mathbf{s}}_N|\mathcal{M})P(\mathcal{M})}{P(\hat{s}_N)} = \frac{P(\hat{\mathbf{s}}_N|\mathcal{M})P(\mathcal{M})}{\sum_{\mathcal{M}'} P(\hat{\mathbf{s}}_N|\mathcal{M}')P(\mathcal{M}')} \\
&= \frac{P(\mathcal{M}) \int d\mathbf{g}_{\mathcal{M}} P(\hat{\mathbf{s}}_N|\mathbf{g}_{\mathcal{M}}, \mathcal{M}) P(\mathbf{g}_{\mathcal{M}}|\mathcal{M})}{\sum_{\mathcal{M}'} P(\mathcal{M}') \int d\mathbf{g}_{\mathcal{M}'} P(\hat{\mathbf{s}}_N|\mathbf{g}_{\mathcal{M}'}, \mathcal{M}') P(\mathbf{g}_{\mathcal{M}'}|\mathcal{M}')} \\
&\propto P(\mathcal{M}) \int d\mathbf{g}_{\mathcal{M}} \prod_i^N \left(e^{\sum_{\mu \in \mathcal{M}} g^\mu \phi^\mu(\mathbf{s}^{(i)})} \right) e^{-N \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}})} P(\mathbf{g}_{\mathcal{M}}|\mathcal{M}) \\
&= P(\mathcal{M}) \int d\mathbf{g}_{\mathcal{M}} e^{N\left(\sum_{\mu \in \mathcal{M}} g^\mu \overline{\phi^\mu}(\hat{\mathbf{s}}_N) - \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}})\right)} P(\mathbf{g}_{\mathcal{M}}|\mathcal{M})
\end{aligned}
\tag{2.14}
$$

where we defined the *empirical averages*: $\overline{\phi^\mu} = \frac{1}{N} \sum_{i=1}^N \phi^\mu(\mathbf{s}^{(i)})$.
We get the saddle point:

$$
\hat{\mathbf{g}}_{\mathcal{M}} = \arg\max_{\mathbf{g}_{\mathcal{M}}} \left(\sum_{\mu \in \mathcal{M}} g^\mu \overline{\phi^\mu}(\hat{\mathbf{s}}_N) - \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}}) \right)
$$

so that for $N \gg 1$:

$$
\begin{aligned}
\mathcal{L}_{\hat{\mathbf{s}}_N}(\mathbf{g}_{\mathcal{M}}) &\equiv \sum_{\mu \in \mathcal{M}} g^\mu \overline{\phi^\mu}(\hat{\mathbf{s}}_N) - \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}}) \\
&= \left(\sum_{\mu \in \mathcal{M}} \hat{g}^\mu \overline{\phi^\mu}(\hat{\mathbf{s}}_N) - \log \mathcal{Z}_{\mathcal{M}}(\hat{\mathbf{g}}_{\mathcal{M}}) \right) + \\
&\quad - \frac{N}{2} \sum_\mu \sum_\nu (g^\mu - \hat{g^\mu})(g^\nu - \hat{g^\nu}) \partial_\mu \partial_\nu \log \mathcal{Z}_{\mathcal{M}}(\hat{\mathbf{g}}_{\mathcal{M}}) + \\
&\quad + o((g - \hat{g})^3) \\
&= \mathcal{L}_{\hat{\mathbf{s}}_N}(\hat{\mathbf{g}}_{\mathcal{M}}) - \frac{N}{2} \sum_\mu \sum_\nu \Delta^\mu J_{\mu\nu}^{[\hat{s}_N]}(\hat{\mathbf{g}_{\mathcal{M}}}) \Delta^\nu + o(\Delta^3)
\end{aligned}
\tag{2.15}
$$

(with $\Delta^\mu = g^\mu - \hat{g}^\mu$) . Notice how the J matrix can here be defined directly in terms of the partition function $\mathcal{Z}_{\mathcal{M}}$, since the second order derivatives of all "single operator" terms in the log likelihood vanish; this means that we can write:

$$
\begin{aligned}
J_{\mu\nu}^{[\hat{s}_N]}(\hat{\mathbf{g}_{\mathcal{M}}}(\hat{s}_N)) &= -\partial_\mu \partial_\nu \overline{\log P(s|\mathbf{g}_{\mathcal{M}}, \mathcal{M})}|_{\hat{\mathbf{g}_{\mathcal{M}}}(\hat{s}_N)} \\
&= \partial_\mu \partial_\nu \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}})|_{\hat{\mathbf{g}_{\mathcal{M}}}(\hat{s}_N)}
\end{aligned}
\tag{2.16}
$$

By the same line of reasoning, we get for the Fisher information:

$$\begin{aligned}\mathbb{J}_{\mu\nu}(\mathbf{g}_{\mathcal{M}}) =& \mathop{\mathbb{E}}_{P(s|\mathbf{g}_{\mathcal{M}},\mathcal{M})}[-\partial_\mu\partial_\nu \log P(s|\mathbf{g}_{\mathcal{M}},\mathcal{M})] \\ =& \mathop{\mathbb{E}}_{P(s|\mathbf{g}_{\mathcal{M}},\mathcal{M})}[\partial_\mu\partial_\nu \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}})] \\ =& \partial_\mu\partial_\nu \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\mathcal{M}})\end{aligned} \tag{2.17}$$

so that if we evaluate the latter for $\mathbf{g}_{\mathcal{M}} = \mathbf{g}_{\hat{\mathcal{M}}}(\hat{s}_N)$, it coincides with the former.
Let's see how this affects the computation: we are now going to insert both the second order loglikelihood expansion and the appropriate Jefferys prior (as discussed in Chapter 1) into our expression for the posterior $P(\mathcal{M}|\hat{s}_N)$, drop any normalization constant, and perform saddle point integration:

$$\begin{aligned}P(\mathcal{M}|\hat{s}_N) \propto& \frac{P(\mathcal{M})e^{N\overline{\mathcal{L}_{\hat{\mathbf{s}}_N}(\theta^*)}}}{\int_{\mathcal{M}}\sqrt{\det \mathbb{J}(\vec{\theta'})}d\vec{\theta'}}\int_{\mathcal{M}} d\theta_{\mathcal{M}}\sqrt{\det \mathbb{J}(\vec{\theta})}e^{-\frac{N}{2}\sum_{\mu,\nu}^{|\mathcal{M}|} d\theta_\mu d\theta_\nu \partial_\mu\partial_\nu \log \mathcal{Z}_{\mathcal{M}}(\mathbf{g}_{\hat{\mathcal{M}}}(\hat{s}_N))} \\ =& P(\mathcal{M})\frac{\sqrt{\det \mathbb{J}(\theta^*)}}{\sqrt{\det J_{\hat{s}_N}(\theta^*)}}\frac{e^{N\overline{\mathcal{L}_{\hat{s}}(\theta^*)}}}{\int_{\mathcal{M}} d\theta_{\mathcal{M}}\sqrt{\det \mathbb{J}(\theta_{\mathcal{M}})}}\left(\frac{2\pi}{N}\right)^{\frac{|\mathcal{M}|}{2}} \\ =& P(\mathcal{M})\frac{\sqrt{\det \mathbb{J}(\theta^*)}}{\sqrt{\det J_{\hat{s}_N}(\theta^*)}}\exp\left[N\overline{\mathcal{L}_{\hat{s}_N}(\theta^*)} - \frac{|\mathcal{M}|}{2}\log\left(\frac{N}{2\pi}\right) - c_{\mathcal{M}}^{BMS}\right]\end{aligned} \tag{2.18}$$

Which coincides with the expression we found in chapter 1, but with the *relative complexity* term left explicitly written. In fact, in this case this term *gives no contribution whatsoever*, since we've just seen that, when evaluated at the maximum likelihood point, $\mathbb{J}(\theta^*) \equiv J_{\hat{s}_N}(\theta^*)$
One can show that this result is valid not only for spin models, but for the whole exponential family. We can write finally:

$$P(\mathcal{M}|\hat{s}_N) \approx P(\mathcal{M})\exp\left[N\overline{\mathcal{L}_{\hat{s}_N}(\theta^*)} - \frac{|\mathcal{M}|}{2}\log\left(\frac{N}{2\pi}\right) - c_{\mathcal{M}}^{BMS}\right] \tag{2.19}$$

In order to extract useful information from this formula, we first need a couple more tools.

2.3 Gauge transformation & loops

Consider an *independent* collection of n operators:

$$\mathcal{I} = \{\phi^{\mu_1}, \phi^{\mu_2}, \cdots, \phi^{\mu_n}\} \tag{2.20}$$

This constitutes, as we said above, a *generating* set of Ω_n.
We can use this collection to generate a group automorphism $\varphi_{\mathcal{I}}$ of Ω_n (or, equivalently, a group automorphism $\sigma_{\mathcal{I}}$ of the state space $\mathcal{S}_n$, seen as a group equipped with the spinwise product composition operator):

$$\phi^{\mu}(\sigma_{\mathcal{I}}(\mathbf{s})) = \prod_{i=1}^{n} \phi^{\mu_i}(\mathbf{s}) = \phi^{\oplus_{i=1}^{n}\mu_i}(\mathbf{s}) = \phi^{\varphi_{\mathcal{I}}(\mu)}(\mathbf{s}) \tag{2.21}$$

We will call this kind of automorphism a *gauge transformation*.
As group automorphisms, gauge transformations leave the identity element unchanged (be it ϕ^0 in the "operator" picture, or $(+,+,\cdots,+)$ in the "state" one)

Notice that gauge transformations can quite freely change the orders of the operators they act on. If we start from, say, a pairwise model, and act upon it with a suitable gauge tranformation, we end up with a model with the same algebraic structure of the first, but composed of operators of potentially any order. It will then be interesting to study how the complexity of a model changes under such transformations, in order to assess one of the fundamental questions we are interested in, namely: *are models with lower-order interactions simpler?*

Drawing conclusions in this sense requires that we first investigate the *partition function* of a model $\mathcal{M}$.

2.3.1 The partition function $\mathcal{Z}_{\mathcal{M}}(\mathbf{g})$

Since the operators ϕ^{μ} can only assume values $\{-1,+1\}$:

$$\begin{aligned} e^{g^{\mu}\phi^{\mu}(s)} &= \cosh(g^{\mu}\phi^{\mu}(s)) + \sinh(g^{\mu}\phi^{\mu}(s)) \\ &= \cosh(g^{\mu}) + \phi^{\mu}(s)\sinh(g^{\mu}) \end{aligned} \tag{2.22}$$

Now if we write the partition function:

$$\begin{aligned}
\mathcal{Z}_{\mathcal{M}}(\mathbf{g}) =& \sum_{\mathbf{s}} \prod_{\mu}^{\mathcal{M}} e^{g^{\mu}\phi^{\mu}(s)} = \sum_{\mathbf{s}} \prod_{\mu}^{\mathcal{M}} (\cosh(g^{\mu}) + \phi^{\mu}(s)\sinh(g^{\mu})) \\
=& \sum_{\mathbf{s}} \prod_{\mu}^{\mathcal{M}} [\cosh(g^{\mu})(1 + \phi(\mathbf{s})^{\mu}\tanh(g^{\mu}))] \\
=& (\cosh(g^{\mu}))^{|\mathcal{M}|} \sum_{\mathbf{s}} \prod_{\mu}^{\mathcal{M}} (1 + \phi(\mathbf{s})^{\mu}\tanh(g^{\mu})) \\
=& \sum_{\mathbf{s}} \prod_{\mu}^{\mathcal{M}} [\cosh(g^{\mu})(1 + \phi(\mathbf{s})^{\mu}\tanh(g^{\mu}))] \\
=& (\cosh(g^{\mu}))^{|\mathcal{M}|} \sum_{\mathbf{s}} \sum_{\mathcal{M}' \subseteq \mathcal{M}} \prod_{\mu}^{\mathcal{M}'} (\tanh(g^{\mu})) \prod_{\nu}^{\mathcal{M}'} \phi(\mathbf{s})^{\nu} \\
=& (\cosh(g^{\mu}))^{|\mathcal{M}|} \sum_{\mathbf{s}} \sum_{\mathcal{M}' \subseteq \mathcal{M}} \phi^{\oplus_{\mathcal{M}'}\mu}(\mathbf{s}) \prod_{\mu}^{\mathcal{M}'} (\tanh(g^{\mu})) \\
=& (\cosh(g^{\mu}))^{|\mathcal{M}|} \sum_{\mathcal{M}' \subseteq \mathcal{M}} \left[\left(\prod_{\mu}^{\mathcal{M}'} (\tanh(g^{\mu})) \right) \sum_{\mathbf{s}} \phi^{\oplus_{\mathcal{M}'}\mu}(\mathbf{s}) \right]
\end{aligned} \tag{2.23}$$

If we recall the orthogonality property of spin operators, we can see that in the sum over models $\mathcal{M}'$ only those terms for which

$$\underset{\mu \in \mathcal{M}'}{\oplus} \mu \equiv 0 \tag{2.24}$$

will give a nonzero contribution (precisely, they'll contribute with a 2^n). We will call subset of operators obeying equation (2.24) *loops*, and denote them from now on with the letter l. The set of all loops contained in a model $\mathcal{M}$ will be denoted by $\mathcal{L}(\mathcal{M})$.
The partition function will then become:

$$\mathcal{Z}_{\mathcal{M}}(\mathbf{g}) = 2^n (\cosh(g^{\mu}))^{|\mathcal{M}|} \sum_{l \in \mathcal{L}(\mathcal{M})} \left[\left(\prod_{\mu}^{l} (\tanh(g^{\mu})) \right) \right] \tag{2.25}$$

Notice how its functional form *depends on the model $\mathcal{M}$ only through*:

- *its "loop structure" $\mathcal{L}(\mathcal{M})$*
- *the total number $|\mathcal{M}|$ of operators in it.*

2.3.2 Invariance of loop structures

Here is the main claim:

The loop structure $\mathcal{L}(\mathcal{M})$ is invariant under gauge transformations.

In order to prove this, let's first check that:

$$\begin{aligned}\phi^{\varphi_{\mathcal{I}}(\mu)}(\mathbf{s})\phi^{\varphi_{\mathcal{I}}(\nu)}(\mathbf{s}) =& \prod_{i\in\mu}\phi^{\rho_i}(\mathbf{s})\prod_{i\in\nu}\phi^{\rho_i}(\mathbf{s})\\ =& \prod_{\{i|i\in\mu\wedge i\in\nu\}}(\phi^{\rho_i}(\mathbf{s}))^2\prod_{\{i|i\in\mu\wedge i\notin\nu\}}\phi^{\rho_i}(\mathbf{s})\prod_{\{i|i\notin\mu\wedge i\in\nu\}}\phi^{\rho_i}(\mathbf{s})\\ =& 1\cdot\prod_{\{i|i\in\mu\wedge i\notin\nu\}}\phi^{\rho_i}(\mathbf{s})\prod_{\{i|i\notin\mu\wedge i\in\nu\}}\phi^{\rho_i}(\mathbf{s}) = \prod_{i\in\mu\oplus\nu}\phi^{\rho_i}(\mathbf{s})\\ =& \phi^{\varphi_{\mathcal{I}}(\mu\oplus\nu)}(\mathbf{s})\end{aligned} \tag{2.26}$$

meaning that $\oplus$ and $\varphi_{\mathcal{I}}$ commute.
It is now immediate to check that:

- a loop l, under a gauge transformation, gets mapped into a loop $l' = \varphi_{\mathcal{I}}(l)$ of the same length;
- a collection $\mathcal{M}'$ of operators which do *not* form a loop gets mapped, under a gauge transformation, into a collection of operators not forming a loop as well.

This means that our expression for $\mathcal{Z}_{\mathcal{M}}(\mathbf{g})$ is completely insensitive to gauge transformations! Now if we reconsider (2.17) we arrive at the necessary conclusion that not only $\mathcal{Z}_{\mathcal{M}}(\mathbf{g})$, but **also $c_{\mathcal{M}}^{BMS}$ is a gauge-invariant quantity**.

This is the main result. We stop here: a lot of mathematical characterization of loops, models, and actual values of the geometric complexity can be brought through: for all this we refer the reader to ([4]).

2.4 So what?

What can we learn from this result? Well, the main point we get to is that *model selection alone does not justify restriction to pairwise models.*
This is because, as we've seen before, gauge transformations can change and shuffle quite freely the orders of interactions of a given model. Now, once we have fixed the number n of spins and the number $|\mathcal{M}|$ of operators to be used, the only quantity left helping us to discriminate between models' posterior probabilities is $c_{\mathcal{M}}^{BMS}$. The fact that this quantity is gauge-invariant means

that the set of all possible models is effectively partitioned in *classes of complexity* which are basically classes of equivalence with respect to gauge tranformations.
Inspection of these classes reveals that lower-order interaction are not a symptom of simplicity, while actual proxies we can look for in this sense exist: in fact, each complexity class contains models with interactions of very different orders.

Higher-order interactions are especially justified in all contexts in which we deal with *latent* (unobserved) variables (see for instance [23], [24]) - and this is usually the case when studying complex systems. For those reasons, our following efforts will address the problem of model selection within the class of spin models with interactions of arbitrary order. In the next chapter we will present and discuss an heuristic to perform such selection without having to enforce any prior restriction on possible models.

Chapter 3

A heuristic for spin model selection

Introduction

3.0.1 Outline of the chapter

The number of possible nondegenerate n-spin models with interactions of arbitrary order is $2^{|\Omega_n|} = 2^{2^n-1}$. The task of selecting a "best candidate" in this set on the basis of N observations (where N, in the undersampling regime we commonly find ourselves in while studying complex systems, can easily be of order $\sim 2^n$) is evidently unfeasible.
Yet it is reasonable to think that the systems we are studying will not be completely disordered; we actually expect that in any interesting system there will be a fair amount of structured dependencies: our aim is to represent these in a statistical model of interacting variables, without assuming the structure of such model a priori.

In this chapter we present an heuristic which should make the task of selecting between so many models less daunting; we do so by a two-step procedure in which:

- We first perform model selection in the space of *mixture models*, where this operation is, as we will see, straightforward;
- We then project the results obtained on the class of spin models.

This procedure has first been described in [6]. We will mainly follow that work, together with [25], for this exposition, adding some original computations and interpretations of results. We close presenting some numerical results and discussing them.

3.0.2 "Why mixtures?" General motivation and the inverse formula for $\hat{g}$

We can derive a useful formula from the *completeness* property of generalized spin systems; if we start from the maximum entropy distribution

$$P(\mathbf{s}|\vec{g}) = e^{\sum_\mu g^\mu \phi^\mu(\mathbf{s})}$$

and take logarithms:

$$\log P(\mathbf{s}|\vec{g}) = \sum_\mu g^\mu \phi^\mu(\mathbf{s}),$$

then multiply both sides by $\phi^\nu(\mathbf{s})$ and sum over $\mathbf{s}$:

$$\sum_{\mathbf{s}} \phi^\nu(\mathbf{s}) \log P(\mathbf{s}|\vec{g}) = \sum_\mu g^\mu \sum_{\mathbf{s}} \phi^\nu(\mathbf{s})\phi^\mu(\mathbf{s})$$

Then thanks to completeness:

$$\sum_{\mathbf{s}} \phi^\nu(\mathbf{s}) \log P(\mathbf{s}|\vec{g}) = \sum_\mu g^\mu 2^n \delta_{\mu\nu} = 2^n g^\nu.$$

We thus have an *inverse formula* for the values of couplings given the probabilities of states:

$$g^\mu = \frac{1}{2^n} \sum_{\mathbf{s}} \phi^\mu(\mathbf{s}) \log P(\mathbf{s}|\vec{g}) \tag{3.1}$$

This is interesting because now if we somehow manage to obtain decent estimates for the values $P(\mathbf{s})$ from repeated observations, we can then immediately get estimates for the strength of interactions, without performing numerical optimizations of any kind.
However, attempting this without the necessary care will result in wild overfitting: after all, we have as many free parameters as many states there are - we are in a situation analogous to the high-degree polynomial "perfect fit" we saw in chapter 1.
We already know the proper "cure" for overfitting is model selection; what this formula suggests is that maybe we can try performing such selection forgetting for a moment about the spin representation and working on *mixtures*, these being models whose parameters are the probabilities of individual states themselves; we thus perform selection within the class of mixture models, and only then map our estimate into the spin representation, following the formula above. Model selection entails dimensionality reduction in mixture space, in the form of enforced simmetries between probabilities $P(\mathbf{s})$ of different states: if we are lucky and do things properly, as we pass through (3.1) to the spin representation it is reasonable to hope for dimensionality reduction to be mantained also in $\vec{g}$ space.

In fact, mapping a selected mixture model onto a spin one will effectively provide definition of what are the *sufficient statistics* of the model; these will come in the form of specific linear combinations $\psi(s) = \sum_\mu u_\mu \phi^\mu(s)$ of spin operators, and will turn out to be way less, in number, than the spin operators themselves (their number will be of the order of the number of different *empirical frequencies* $\hat{k}_s$ with which states occur in our dataset). This way we will achieve a data-driven separation between *relevant* and irrelevant variables.

This being the logical path we have to keep in mind, let's now see things in detail.

3.1 Selection on mixtures: fundamentals

3.1.1 The danger of over-resolving states

Consider a system composed of n binary spins, of which only k are interacting. The Hamiltonian of such system will look like this:

$$\mathcal{H}_0(\mathbf{s}) = \sum_{\mu \in \mathcal{M}_{[k]}} g^\mu \phi^\mu(\mathbf{s}) \tag{3.2}$$

where $\mathcal{M}_{[k]}$ denotes a model composed of operators which only involve our chosen k spins.
We expect the marginal distributions for the states of the *noninteracting* units to be uniform:

$$\begin{aligned} P(s_{k+1} = \sigma_{k+1}, s_{k+2} = \sigma_{k+2}, \cdots, s_n = \sigma_n) &= \frac{1}{\mathcal{Z}} \sum_{s_1,\cdots,s_k} e^{-\beta \mathcal{H}(s_1,\cdots,s_k,\sigma_{k+1},\cdots,\sigma_n)} \\ &= \frac{1}{2^{n-k}} \frac{\sum_{s_1,\cdots,s_k} e^{-\beta \mathcal{H}(s_1,\cdots,s_k,\sigma_{k+1},\cdots,\sigma_n)}}{\sum_{s_1,\cdots,s_k} e^{-\beta \mathcal{H}(s_1,\cdots,s_k,\sigma_{k+1},\cdots,\sigma_n)}} \\ &= \frac{1}{2^{n-k}} \end{aligned} \tag{3.3}$$

Now imagine we find ourselves in the position of who knows nothing about the system under investigation, and tries to perform inference from scratch on the n spin assembly.
What we would consider in this case a "good" protocol of inference is one that leads us to recognizing that $n - k$ of our spins are *free*: we would like to end up with a selected model not containing operators which are not in $\mathcal{H}_0(\mathbf{s})$.
First, let us restate the by now established fact that maximum likelihood alone does not yield satisfying results in this context: in fact, when we apply

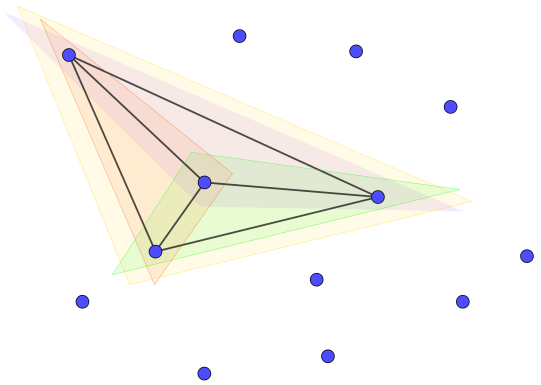

Figure 3.1: *A subcomplete spin system*
Only k of the n spins in our model are actually interacting. Since we have interactions of arbitrary order, pictorial representation of our spin system requires we resort to a *hypergraph* - a simple graph being able to account only for singlewise and pairwise interactions (links).

maximum likelihood to obtain estimates for the generative probabilities of states from sample statistics, we get:

$$\begin{aligned} P(s)_{MLE} &= \arg \max_{\hat{\rho}, \sum \rho_r = 1} P\left(\hat{s}_N \mid \vec{\rho}\right) \stackrel{def}{\equiv} \arg \max_{\hat{\rho}, \sum \rho_r = 1} P\left(\hat{s}_N \mid (P(r) = \rho_r \ \forall r)\right) \\ &= \arg \max_{\hat{\rho}, \sum \rho_r = 1} \prod_{i=1}^{N} \rho_{s^{(i)}} \\ &= \arg \max_{\hat{\rho}, \sum \rho_r = 1} \prod_{s} \rho_s^{\hat{k}_s} = \arg \max_{\hat{\rho}, \sum \rho_r = 1} \log \left(\prod_{s} \rho_s^{\hat{k}_s} \right) \\ &= \arg \max_{\hat{\rho}, \sum \rho_r = 1} \sum_{s} \hat{k}_s \log \rho_s \end{aligned}$$

where $\hat{k}_s$ is the *sample frequency* of state s, i.e. the number of times this state is observed in the dataset. This equation is solved by enforcing:

$$\delta_\rho(\sum_s \hat{k}_s \log \rho_s) - \alpha \delta_\rho(\sum_s \rho_s) = 0$$

leading to

$$P(s)_{MLE} = \rho_s^* = \frac{\hat{k}_s}{N} \tag{3.4}$$

This result implies that our maximum likelihood estimates for the probabilities of states will spot *equiprobable* ones as actually being so *only if we observe all such states the same exact amount of times* - the probability of this event occurring being way more than negligible in all practical situations. In all other cases, we end up with a distribution that *fits the noise*, giving us all sorts of nonzero spurious couplings when we pass to the spin representation.

3.1.2 A two-states example

Assume we are in the simple case $n = 1$: we have 1 binary unit and a series of N independent observations of its state. There are two possible mixture models:

$$\begin{aligned} \mathcal{M}_0 &= \{\text{“}P(1) = P(-1) = 1/2''\} \\ \mathcal{M}_1 &= \{P(1) = \rho;\ P(-1) = 1 - \rho\} \end{aligned} \tag{3.5}$$

Geometrically, the first contitutes a point and the second a one-dimensional curve in the space of distributions.
Let's now compute the respective posterior probabilities:

$$\begin{aligned} P(\mathcal{M}_0|\hat{s}_N) =& \frac{P(\hat{s}_N|\mathcal{M}_0)P(\mathcal{M}_0)}{P(\hat{s}_N)} = \frac{1}{2^N}\frac{P(\mathcal{M}_0)}{P(\hat{s}_N)} \\ \\ P(\mathcal{M}_1|\hat{s}_N) =& \frac{P(\hat{s}_N|\mathcal{M}_1)P(\mathcal{M}_1)}{P(\hat{s}_N)} = \frac{P(\mathcal{M}_1)\int_\rho d\rho P(\hat{s}_N|\rho,\mathcal{M}_1)P(\rho|\mathcal{M}_1)}{P(\hat{s}_N)} \\ =& \frac{P(\mathcal{M}_1)\int_\rho d\rho \rho^{\hat{k}_1}(1-\rho)^{N-\hat{k}_1}P(\rho|\mathcal{M}_1)}{P(\hat{s}_N)} \end{aligned} \tag{3.6}$$

Now we need to fix a prior $P(\rho|\mathcal{M})$ on the distribution of the parameter $\rho \in [0, 1]$. A convenient choice for mixture models are *Dirichlet priors*:

$$P(\rho|\mathcal{M})d\rho = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)}\rho^{a_1-1}(1-\rho)^{a_2-1}d\rho \tag{3.7}$$

They are conjugate priors with respect to multinomial (mixture) sampling distributions[1]. Plus, they are quite expressive priors, this meaning that for suitable values of their parameters they can represent in an efficient manner a more than sufficient amount of possible states of knowledge [2].
We will choose $a_1 = a_2 = a$ since we have no prior information justifying

[1] This basically means that, using these, posteriors will mantain the same functional form as priors, which is really useful in case of repeated updatings of our state of knowledge via subsequent observations.

[2] Still, see [26] for some criticism of these priors.

asymmetry between the probabilities of the two possible outcomes. Inserting the chosen prior into our expression, we find:

$$\begin{aligned} P(\mathcal{M}_1|\hat{s}_N) =& \frac{P(\mathcal{M}_1)}{P(\hat{s}_N)} \frac{\Gamma(2a)}{\Gamma(a)^2} \int_\rho d\rho \rho^{a-1+\hat{k}_1} (1-\rho)^{a-1+N-\hat{k}_1} \\ =& \frac{P(\mathcal{M}_1)}{P(\hat{s}_N)} \frac{\Gamma(2a)}{\Gamma(a)^2} \frac{\Gamma(a+\hat{k}_1)\Gamma(a+N-\hat{k}_1)}{\Gamma(2a+N)} \end{aligned} \tag{3.8}$$

We can now evaluate the ratio between posteriors:

$$\frac{P(\mathcal{M}_1|\hat{s}_N)}{P(\mathcal{M}_0|\hat{s}_N)} = \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} 2^N \frac{\Gamma(2a)}{\Gamma(a)^2} \frac{\Gamma(a+\hat{k}_1)\Gamma(a+N-\hat{k}_1)}{\Gamma(2a+N)} \tag{3.9}$$

We will work with the "uniform" prior $a = 1$; for a detailed discussion of this and other possible choices, see Appendix A.

Model $\mathcal{M}_1$ is to be preferred over $\mathcal{M}_0$ if:

$$1 < \frac{P(\mathcal{M}_1|\hat{s}_N)}{P(\mathcal{M}_0|\hat{s}_N)} = \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} \frac{1}{N+1} \left[\frac{1}{2^N} \binom{N}{\hat{k}_1} \right]^{-1} \tag{3.10}$$

We can take logarithms, and evaluate factorials via Stirling's approximation (assuming $N >> 1$):

$$\begin{aligned} 0 <& \log \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} - \log(N+1) - \log \binom{N}{\hat{k}_1} + N \log 2 \\ \approx& \log \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} - NH\left[\hat{k}_1/N\right] + \log \sqrt{\frac{2\pi}{N}} + \frac{1}{2} \left(\log \frac{\hat{k}_1}{N} + \log(1 - \frac{\hat{k}_1}{N}) \right) \end{aligned}$$

Expanding around the maximum $x_1 \equiv \frac{\hat{k}_1}{N}$ and truncating at second order in $(x_1 - \frac{1}{2})$:

$$0 < \log \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} - \log 2 + \log \sqrt{\frac{2\pi}{N}} + 2 \left(x_1 - \frac{1}{2} \right)^2 (N-1)$$

which finally leads to:

$$\left| \frac{\hat{k}_1}{N} - \frac{1}{2} \right| > \sqrt{\frac{\log \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} + \log \sqrt{\frac{2N}{\pi}}}{2N}} \sim \sqrt{\frac{\log \sqrt{\frac{2N}{\pi}}}{2N}} \tag{3.11}$$

So we see that Bayesian analysis returns the desired answer: a model in which the two states are to be considered separate is to be preferred only if the difference in the frequencies of observations between the two is enough to be *statistically significant*; *relative* fluctuations of frequencies are more and more significant the more sample points we draw, in the spirit of the weak law of large numbers.

We show in figure 3.2 a generalization to the case $n = 3$.

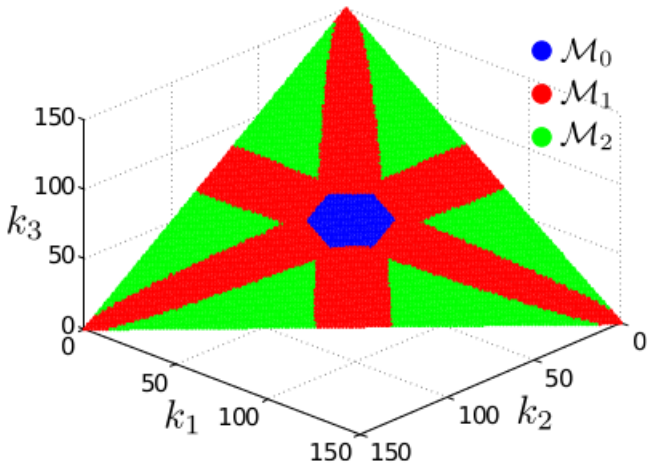

Figure 3.2: *Acceptance regions for different models*
Partition of the 3-simplex in sets of empirical distributions corresponding to different "best" models. Blue: "single-state" model $\mathcal{M}_0$ is preferred; Red: a model clamping together two of the three states is preferred; Green: the model $\mathcal{M}_f$ resolving all three states is preferred. *Image from: A. Haimovici, M. Marsili, "Criticality of most informative samples: a Bayesian model selection approach"* ([25])

What did we just learn?

The main conceptual result is a quantitative restatement of a trivial assert: in absence of particular prior information, *we are not allowed, in general, to consider states as if they were different (for what concerns their probability) if we see them a sufficiently close number of times.*
"The definition of states s is made by the observer, not by the system" [25]. Configuration of planets has no influence on a chemistry experiment, yet sampling noise could well make planets look as relevant variables! The fact that Bayesian analysis gives a quantitative tool for spotting invariances in our data is thus more and more valuable as the systems we look in become more complex, and our intuitive understanding of their inner functioning decreases.

3.2 Selection on mixtures: the general case

Let's now study in detail the general case, in which we have an arbitrary number of states. Here a mixture model formally corresponds to a particular *partition*

$$\mathcal{Q} = \{Q_1, Q_2, \cdots, Q_{|\mathcal{Q}|}\}$$

$$Q_i \bigcap Q_j = \emptyset, \qquad 1 \leq i < j \leq |\mathcal{Q}|$$

$$\bigcup_i Q_i = \mathcal{S}$$

of the space $\mathcal{S}$ of states. Selection will thus be performed among all possible partitions $\mathcal{Q}$.
There is no reason to expect that the behavior of the posteriors observed in the 2-states case would change by raising the number of possible states (and indeed, it doesn't, as has been checked in [25]). Thus, in general, the *frequency partition* $\mathcal{K}$, i.e. the partitioning of states clamping together the ones observed *exactly* the same number of times, has a posterior probability which is higher than the finest one $\mathcal{S}$. This is not enough: to understand "how good is $\mathcal{K}$" we need full characterization of the posterior distribution $P(\mathcal{Q}|\hat{s}_N)$.

Notation

We will refer to partition sets Q_j as $\mathcal{Q}$-states (groups of "equivalent" **s**-states). Their cardinalities will be denoted by $m_j = |Q_j|$.
The parameters of the mixture model are the probabilities "ρ_j" of Q-states:

$$P(\mathbf{s} \in Q_j) = \rho_j \tag{3.12}$$

$\vec{\rho}_{\mathcal{Q}}$ as vectors belong to the $(|\mathcal{Q}| - 1)$ - dimensional probability simplex[3] in $\mathbb{R}^{|\mathcal{Q}|}$.
States within the same partition set have the same probability. We denote these by:

$$P(\mathbf{s} \mid \mathbf{s} \in Q_j) = \mu_j = \frac{\rho_j}{m_j} \tag{3.13}$$

The normalization for $\vec{\mu}$ thus becomes:

$$\sum_j m_j \rho_j = 1 \tag{3.14}$$

3.2.1 Computation of model posteriors $P(\mathcal{Q}|\hat{s}_N)$

The posterior probability of a partition $\mathcal{Q}$ reads:

$$P(\mathcal{Q}|\hat{s_N}) = \frac{P(\mathcal{Q})P(\hat{s}_N|\mathcal{Q})}{P(\hat{s}_N)} \tag{3.15}$$

If we call $\mathcal{Q}_0$ the trivial partition composed of a single set containing all states, we can expand $\forall \mathcal{Q} \neq \mathcal{Q}_0$:

$$P(\mathcal{Q}|\hat{s_N}) = \frac{P(\mathcal{Q})}{P(\hat{s}_N)} \int_{\vec{\rho}_{\mathcal{Q}}} d\vec{\rho}_{\mathcal{Q}} P(\hat{s}_N|\vec{\rho}_{\mathcal{Q}}\mathcal{Q}) P(\vec{\rho}_{\mathcal{Q}}|\mathcal{Q}) \tag{3.16}$$

[3]This is defined by the constraints: $\sum_j \rho_j = 1$, $\rho_j > 0 \; \forall j < |\mathcal{Q}|$.

Inside the integral in the right hand side, we must insert the *likelihood*:

$$P(\hat{s}_N|\vec{\rho}_{\mathcal{Q}},\mathcal{Q}) = \prod_{i=1}^{N} \rho_{\mathcal{Q}j(s^{(i)})} = \prod_{j}^{\mathcal{Q}} (\rho_{\mathcal{Q}j})^{K_j(\hat{s}_N)} \tag{3.17}$$

(where $j(s^{(i)})$ as an index identifies the partition set Q_j to which the state s belongs, and $K_j(\hat{s}_N) = \sum_{s\in Q_j} \hat{k}_s$ is the total number of observations of states belonging to the partition set Q_j); and we must insert the *prior*, for which we choose symmetric Dirichlet functions for each $\mathcal{Q}$:

$$P(\vec{\rho}_{\mathcal{Q}}|\mathcal{Q}) = \frac{\Gamma(|\mathcal{Q}|a_{\mathcal{Q}})}{\Gamma(a_{\mathcal{Q}})^{|\mathcal{Q}|}} \prod_{j}^{\mathcal{Q}} \rho_{\mathcal{Q}j}^{a_{\mathcal{Q}}-1} \delta\big(\sum_j \rho_{\mathcal{Q}j} - 1\big) \tag{3.18}$$

Whith these two objects in place, we obtain:

$$P(\mathcal{Q}|\hat{s}_N) = \int_{\vec{\rho}_{\mathcal{Q}}} d\vec{\rho}_{\mathcal{Q}} \frac{P_0(\mathcal{Q})}{P_0(\hat{s_N})} \frac{\Gamma(a_Q Q)}{\Gamma(a_{\mathcal{Q}})^Q \Gamma(a_{\mathcal{Q}} Q + N)} \prod_{j=1}^{Q} \left[\frac{\Gamma(a_{\mathcal{Q}} + K_j)}{m_j^{K_j}}\right] \tag{3.19}$$

Finally, we choose $a_{\mathcal{Q}} = 1\ \forall \mathcal{Q}$ (see Appendix A for discussion of prior choices), and find:

$$P(\mathcal{Q}|\hat{s_N}) = \frac{P_0(\mathcal{Q})}{P_0(\hat{s_N})} \frac{(|\mathcal{Q}|-1)!}{(N+|\mathcal{Q}|-1)!} \prod_{j=1}^{|\mathcal{Q}|} \left[\frac{K_j!}{m_j^{K_j}}\right] \tag{3.20}$$

3.2.2 The optimal partition $\mathcal{Q}^*$

A "best candidate" partition is now model $\mathcal{Q}^*$ maximizing the posterior just obtained:

$$\mathcal{Q}^* = \arg\max_{\mathcal{Q}} P(\mathcal{Q}|\hat{s}_N) \tag{3.21}$$

The numerical task of finding this partition has been studied and discussed in detail in [25]. For our purposes, the main insight we need to borrow from that work is the fact that $\mathcal{Q}^*$ *appears to be always a coarse-graining of* $\mathcal{K}$, i.e. a partition obtained by just clamping together specific (precisely: "adjacent" in frequency, as one would expect) sets of the latter;

We thus end up, after little effort, with *two* possible choices for the $P(\mathbf{s})$ estimate to be inserted in 3.1, namely:

$$P(\mathbf{s}|\hat{s}_N,\mathcal{K}) = \mu_{\mathcal{K}j(s)} = \frac{\rho_{\mathcal{K}j(s)}}{m_{\mathcal{K}j(s)}}$$

and

$$P(\mathbf{s}|\hat{s}_N,\mathcal{Q}^*) = \mu_{\mathcal{Q}^*j(s)} = \frac{\rho_{\mathcal{Q}^*j(s)}}{m_{\mathcal{Q}^*j(s)}}$$

In which the vectors $\vec{\rho}_{\mathcal{Q}}$ are to be regarded as *random* vectors distributed as $P(\vec{\rho}_{\mathcal{Q}}|\hat{s}_N, \mathcal{Q})$. The obtained $\vec{g}$ couplings will be consequently treated as random variables themselves.

3.3 Projection on spin models: fundamentals

3.3.1 Symmetries and dimensionality reduction

Regardless of how our estimates $\rho_{\mathcal{Q}j(s)}$ turn out to be actually distributed, what is really important here is that they will in any case depend on the sample $\hat{s}_N$ *only through* frequencies of observations (remember: $\mathcal{Q}^*$ is a coarse-graining of $\mathcal{K}$). This assures that we can in any case rewrite:

$$\begin{aligned} g^\mu =& \frac{1}{2^n}\sum_{\mathbf{s}}^{\mathcal{S}} \phi^\mu(\mathbf{s}) \log \mu(\hat{k}(s)) = \frac{1}{2^n}\sum_k \left(\sum_{s,\ \hat{k}(s)=k} \phi^\mu(s) \right) \log \mu(k) \\ =& \frac{1}{2^n}\sum_{j=1}^{\mathcal{K}} \chi_j^\mu \log \mu_j \end{aligned} \tag{3.22}$$

where last sum is over the sets $\mathcal{K}_j \in \mathcal{K}$, and $\chi_j^\mu = \sum_{s\in\mathcal{K}_j} \phi^\mu(s)$. Notice how the final formula returns the components of a $(2^n - 1)$ - dimensional vector as linear functions of a $|\mathcal{K}| - 1$ - dimensional one (both "-1" being due to normalization). This means that the vector $\vec{g}$ must necessarily lie on a $|\mathcal{K}| - 1$ dimensional surface in the g-space, and this in turn means that it should in principle be possible to redefine our basis of operators $\{\phi^\mu\}$ in order to obtain a model which only contains $(|\mathcal{K}| - 1)$ operators. This would produce a model which is *sparse* in this new representation.

Example: full-pairwise generative model

We closely follow the exposition in [6].
Consider a four-spin system with a Ising-like generative Hamiltonian, i.e. an Hamiltonian comprising all possible two-spin interactions (with a shared coupling constant J) and no interaction of any other order:

$$\mathcal{H}_0(s) = J\left(\phi^{(1,2)}(s) + \phi^{(1,3)}(s) + \phi^{(1,4)}(s) + \phi^{(2,3)}(s) + \phi^{(2,4)}(s) + \phi^{(3,4)}(s)\right) \tag{3.23}$$

We give a pictorial representation of this system in figure 3.3. The model is *exchangeable*, i.e. insensitive to permutations of the n spins, and it is not hard to convice oneself that due to degeneracy of parameters, the set of states is naturally partitioned into three classes of equiprobable ones; in fact, given the low dimensionality we can check by hand:

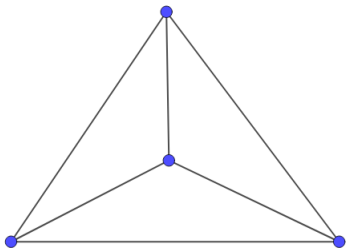

Figure 3.3: Full-pairwise 4-spin model.

s	$\mathcal{H}(s)$	s	$\mathcal{H}(s)$	s	$\mathcal{H}(s)$	s	$\mathcal{H}(s)$
$++++$	$6J$	$+-++$	0	$-+++$	0	$--++$	$-2J$
$+++-$	0	$+-+-$	$-2J$	$-++-$	$-2J$	$--+-$	0
$++-+$	0	$+--+$	$-2J$	$-+-+$	$-2J$	$---+$	0
$++--$	$-2J$	$+---$	0	$-+--$	0	$----$	$6J$

So that we get a "true partitioning" $\mathcal{Q}_{gen}$ composed of the three sets:

$$Q_j = \{s : s_1 + s_2 + s_3 + s_4 = (+2J, 0, -2J)\}$$

What we expect is that, with enough data, the optimal partition $\mathcal{Q}^*$ would itself group together members of these three classes; if this is the case, we can check that our formula 3.1 sets automatically to 0 all g^μs corresponding to odd-order (singlewise and triplewise) interactions:

$$g^{(1)} = g^{(2)} = g^{(3)} = g^{(4)} = g^{(1,2,3)} = g^{(1,2,4)} = g^{(1,3,4)} = g^{(2,3,4)} = 0$$

The same formula also outputs nonzero, *all equal* pairwise couplings:

$$g^{(1,2)} = g^{(1,3)} = g^{(1,4)} = g^{(2,3)} = g^{(2,4)} = g^{(3,4)} = \hat{j}(\vec{\rho}^*_{\mathcal{Q}})$$

and a nonzero four-body coupling:

$$g^{(1,2,3,4)} = \hat{c}_4(\vec{\rho}^*_{\mathcal{Q}})$$

In these relations *only the numerical values of* $\hat{j}$ *and* $\hat{c}_4$ *depend on the posterior parameter estimates* $\vec{\rho}_{1\mathcal{Q}^*}$ - the constraints obtained depend on the choice of the partition alone! Equation 3.1 is thus translating the dimensionality reduction obtained via model selection in mixture space into a (even stronger!) dimensionality reduction in spin model space: out of $2^{2^n-1} = 32768$ possible models of four spins, symmetries in data here allowed us to reduce the set of possible models to just *three*:

$$\begin{aligned}
\mathcal{M}_1 &= \{\{\phi\}_{pairwise}, \phi^{(1,2,3,4)}\} \\
\mathcal{M}_2 &= \{\{\phi\}_{pairwise}\} \\
\mathcal{M}_3 &= \{\phi^{(1,2,3,4)}\}
\end{aligned} \tag{3.24}$$

We include the latter two because in principle we could discover that, besides the information contained in the choice of $\mathcal{Q}^*$, the very posterior estimates of the probabilities $\vec{\rho}_{\mathcal{Q}}$ are such that $\hat{j}(\vec{\rho}_{\mathcal{Q}^*}) = 0$ or $\hat{c}_4(\vec{\rho}_{\mathcal{Q}^*}) = 0$. If we check manually, via (3.1), the conditions on single state probabilities under which $\mathcal{M}_2$ or $\mathcal{M}_3$ would be the correct ones. We find that:

- $\mathcal{M}_3$ would correspond to the additional constraint $p_{++++} \equiv p_{++--}$; but, were this true, we would have obtained via mixture selection a further coarse-grained $\mathcal{Q}^*$, not distinguishing between these two states either. The fact that we didn't allows us to rule out this model.

- $\mathcal{M}_2$ corresponds to the additional constraint: $p_{++++}p^3_{+++-} \equiv p^4_{++--}$; now, this condition is *completely transparent* to the eyes of mixture selection; thus, we cannot reject this model on the basis of the obtained $\mathcal{Q}^*$ alone.

We can picture the result of inference as having identified a 2-dimensional manifold in a 2^n-1 - dimensional space; the retrieved manifold is parametrized by $\hat{j}, \hat{c}_4$.
This picture traslates naturally to the general case - we will now see in detail how it does.

3.4 Projection on spin models: the general case

3.4.1 Singular Value Decomposition of χ

Our task will be now the one of characterizing the $(|\mathcal{Q}| - 1)$ - dimensional manifold in $\vec{g}$ space on which our inferred couplings $g^{\mu}_{\mathcal{Q}}$ live.
The natural way to do this is to decompose χ via Singular Value Decomposition (SVD):

$$\begin{aligned} \chi_{\mu j} &= \sum_{\nu \in \Omega_n} \sum_{k \in \mathcal{Q}} U_{\mu\nu} \Lambda_{\nu k} W_{kj} \qquad\qquad (\Lambda_{\nu k} = \Lambda_k \delta_{\nu k}) \\ &= \sum_{k \in \mathcal{Q}} U_{\mu k} \Lambda_k W_{kj} \end{aligned} \tag{3.25}$$

where U is a $2^n - 1 \times 2^n - 1$ unitary matrix (we discard g^0, this being fixed by normalization and thus not being a free parameter of the model) and W is $|\mathcal{Q}| \times |\mathcal{Q}|$ and unitary as well. U in particular allows for changes of basis in the $\vec{g}$ space. If we define a new operator basis:

$$\psi^{\eta}(s) = \sum_{\mu \in \Omega_n} \phi^{\mu}(s) U_{\mu\eta} \tag{3.26}$$

we see that the Hamiltonian can be rewritten:

$$\begin{aligned}\mathcal{H}(s) =& \sum_{\mu} g^{\mu}\phi^{\mu}(s) = \sum_{\mu,\nu} g^{\mu}\delta_{\mu\nu}\phi^{\nu}(s) = \sum_{\mu,\nu} g^{\mu}\left(\sum_{\eta} U_{\mu\eta}U_{\nu\eta}\right)\phi^{\nu}(s) \\ =& \sum_{\eta}\left(\sum_{\mu} g^{\mu}U_{\mu\eta}\right)\left(\sum_{\nu}\phi^{\nu}(s)U_{\nu\eta}\right) = \sum_{\eta}\left(\sum_{\mu} g^{\mu}U_{\mu\eta}\right)\psi^{\eta}(s) \\ =& \sum_{\eta}\tilde{g}^{\eta}\psi^{\eta}(s)\end{aligned}$$

under suitable redefinition of the coupling constants:

$$\tilde{g}^{\eta} = \sum_{\mu} g^{\mu}U_{\mu\eta} \tag{3.27}$$

We now see that *there are at most* $|\mathcal{Q}| - 1$ *nonzero couplings* $\tilde{g}^{\eta}$ in the new representation: inserting (3.25) into (3.1) :

$$g^{\mu} = \frac{1}{2^n}\sum_{j\in\mathcal{Q}}\sum_{k\in\mathcal{Q}} U_{\mu k}\Lambda_k W_{kj}\log\frac{\rho_j}{m_j} \tag{3.28}$$

then inserting this in 3.27:

$$\begin{aligned}\tilde{g}^{\eta} =& \sum_{\mu} g^{\mu}U_{\mu\eta} = \frac{1}{2^n}\sum_{j\in\mathcal{Q}}\sum_{k\in\mathcal{Q}}\left(\sum_{\mu} U_{\mu\eta}U_{\mu k}\right)\Lambda_k W_{kj}\log\frac{\rho_j}{m_j} \\ =& \begin{cases}\dfrac{\Lambda_\eta}{2^n}\displaystyle\sum_{j\in\mathcal{Q}} W_{\eta j}\log\frac{\rho_j}{m_j}, & 1 \le \eta \le |\mathcal{Q}| - 1 \\ \\ 0 & \text{otherwise.}\end{cases}\end{aligned} \tag{3.29}$$

The "-1" comes from the fact that the $\sum_{\mu}\chi_{\mu j} = 0$ constraint translates in a singular value Λ_0 being $= 0$.

3.4.2 Sufficient statistics

If we rewrite the probability distribution for the states in this new basis:

$$P(s|\tilde{g}) = \frac{1}{Z}e^{\sum_{\eta}\tilde{g}_{\eta}\psi^{\eta}(s)} \tag{3.30}$$

we see that these relatively few functions (less than $|\mathcal{K}|$ in any case!) constitute the *sufficient statistics* of the model. [4]. They emerged via a Bayesian

[4]For a mathematical characterization of sufficient statistics, see [27]; for a characterization of their importance in statistical physics, see [28]

model selection procedure, this meaning that they should be robust against overfitting. *We achieved great dimensional reduction*: in the undersampling regime one always has $|\mathcal{K}| \ll 2^n$, and the number of our sufficient statistics is modulated by the former quantity.

The derivation of the statistics ψ^η is the main result of [5], and of the pèresent chapter.
Once the ψ functions are determined, parameter estimation is performed in practical cases by maximum likelihood estimation (see Appendix B). The authors argue in [5] that inference can be made more stable by neglecting the statistics corresponding to the smallest singular values Λ_η. In chapter 4 we will elaborate on this in full detail, by analytically characterizing the spectral properties of the χ matrix in relation to its defining partition $\mathcal{Q}$.

The following section reviews two strategies for the $\vec{g}$ estimation stage; the effort is aimed at gaining some insight about stability of inference.
It can be skipped at a first reading.

3.5 Insights *via* parameter estimation

We'll work with $\mathcal{K}$, but extension to $\mathcal{Q}^*$ requires mere substitutions in the expressions. Our core formula reads:

$$g^\mu_{\mathcal{K}} = \frac{1}{2^n} \sum_j \chi^\mu_j \log \frac{\vec{\rho}_{\mathcal{K}j}}{m_{\mathcal{K}j}} \tag{3.31}$$

Where, we recall, $\vec{\rho}_{\mathcal{K}} \sim P_1(\vec{\rho}_{\mathcal{K}}|\hat{s}_N, \mathcal{K})$ are random vectors, distributed as their $\mathcal{K}$-fixed-model posterior given the sample.

Treating $\vec{g}$ as random vector: stability of inference

A first choice is to accordingly treat $\vec{g}$ as a random variable itself, and try and characterize its distribution. For this, we will need to explicitly write the expression for $P(\vec{\rho}_{\mathcal{K}}|\hat{s}_N, \mathcal{K})$:

$$\begin{aligned} P(\vec{\rho}_{\mathcal{K}}|\hat{s}_N, \mathcal{K}) =& \frac{P(\hat{s}_N|\vec{\rho}_{\mathcal{K}}, \mathcal{K}) P(\vec{\rho}_{\mathcal{K}}|\mathcal{K})}{P(\hat{s}_N)} = \prod_r^{\mathcal{S}} \left(\frac{\rho_{j(r)}}{m_{j(r)}} \right)^{\hat{k}_r} \frac{P(\vec{\rho}_{\mathcal{K}}|\mathcal{K})}{P(\hat{s}_N)} \\ =& \frac{1}{P(\hat{s}_N)} \left[\prod_j^{\mathcal{K}} \left(\frac{\rho_j}{m_j} \right)^{\hat{K}_j} \right] \frac{\Gamma(a|\mathcal{K}|)}{\Gamma(a)^{|\mathcal{K}|}} \left[\prod_j^{\mathcal{K}} (\rho_j)^{a-1} \right] \\ =& C \cdot \left[\prod_j^{\mathcal{K}} (\rho_j)^{\hat{K}_j + a - 1} \right] \end{aligned} \tag{3.32}$$

C being a constant. Normalization now entails:

$$P(\vec{\rho}_{\mathcal{K}}|\hat{s}_N, \mathcal{K}) = \frac{\prod_j \Gamma(\hat{K}_j + a)}{\Gamma(N + a|\mathcal{K}|)} \prod_j^{\mathcal{K}} \rho_j^{\hat{K}_j + a - 1} \; \delta_{(\sum_j \rho_j - 1)} \tag{3.33}$$

Now that we have a full expression of the distribution for $\vec{\rho}$, let's turn to the spin representation and first compute the average couplings (and also drop partition indices):

$$\mathbb{E}\left[g^\mu\right] = \frac{1}{2^n} \sum_j \chi_j^\mu (\mathbb{E}\left[\log \vec{\rho}_j\right] - \log m_j) \tag{3.34}$$

In order to be able to evaluate $\mathbb{E}\left[\log \vec{\rho}_j\right]$ we can define the auxiliary function:

$$Z(\vec{\lambda}) = \mathbb{E}\left[\prod_j \rho_j^{\lambda_j}\right] \tag{3.35}$$

and take derivatives:

$$\partial_{\lambda_j} \log Z(\vec{\lambda})\Big|_{\vec{\lambda} \equiv 0} = \frac{\partial_{\lambda_j} Z(\vec{\lambda})\big|_{\vec{\lambda} \equiv 0}}{Z(0)} = \mathbb{E}\left[\log \rho_j \prod_k \rho_k^{\lambda_k}\Big|_{\vec{\lambda} \equiv 0}\right] = \mathbb{E}\left[\log \rho_j\right] \tag{3.36}$$

Now all we need is an explicit expression for $Z(\vec{\lambda})$:

$$\begin{aligned} Z(\vec{\lambda}) &= \int_{\vec{\rho}} d\vec{\rho} \prod_j \rho_k^{\lambda_j} P(\vec{\rho}|\hat{s}_N, \mathcal{K}) \\ &= \frac{\prod_j \Gamma(\hat{K}_j + a)}{\Gamma(N + a|\mathcal{K}|)} \int_{\vec{\rho}} d\vec{\rho} \prod_j \rho_k^{\lambda_j + \hat{K}_j + a - 1} \; \delta_{(\sum_j \rho_j - 1)} \\ &= \frac{\Gamma(N + a|\mathcal{K}| + \sum_j \lambda_j)}{\Gamma(N + a|\mathcal{K}|)} \prod_j \frac{\Gamma(\hat{K}_j + a)}{\Gamma(\hat{K}_j \lambda_j + a)} \end{aligned} \tag{3.37}$$

so that:

$$\partial_{\lambda_j} \log Z(\vec{\lambda})\Big|_{\vec{\lambda} = 0} = \psi^{(0)}(\hat{K}_j + a) - \psi^{(0)}(N + a|\mathcal{K}|) \tag{3.38}$$

in which $\psi^{(r)}(x) = \frac{d^r}{dx^r} \log \Gamma(x)$ is the polygamma function.
If we now insert this into 3.34 we get:

$$\mathbb{E}\left[g^\mu\right] = \frac{1}{2^n} \sum_j \chi_j^\mu \left(\psi^{(0)}(N + a|\mathcal{K}|) - \psi^{(0)}(\hat{K}_j + a) - \log m_j\right) \tag{3.39}$$

We don't need to further characterize this object here. What we care much more about is the *covariance*:

$$\begin{aligned} Cov\left[g^{\mu}g^{\nu}\right] =& \frac{1}{2^{2n}}\sum_{j,k}\chi_j^{\mu}\chi_k^{\nu}Cov\left[(\log\rho_j - \log m_j)(\log\rho_k - \log m_k)\right] \\ =& \frac{1}{2^{2n}}\sum_{j,k}\chi_j^{\mu}\chi_k^{\nu}Cov\left[\log\rho_j \log\rho_k\right] \end{aligned} \tag{3.40}$$

We again exploit $Z(\vec{\lambda})$; note:

$$\begin{aligned} \partial_{\lambda_j}\partial_{\lambda_k}\log Z(\vec{\lambda})\Big|_{\vec{\lambda}=0} =& \frac{\partial_{\lambda_j}\partial_{\lambda_k}Z(\vec{\lambda})\big|_{\vec{\lambda}=0} - \partial_{\lambda_j}Z(\vec{\lambda})\big|_{\vec{\lambda}=0}\partial_{\lambda_k}Z(\vec{\lambda})\big|_{\vec{\lambda}=0}}{Z(0)} \\ =& \mathbb{E}\left[\log\rho_j\log\rho_k\right] - \mathbb{E}\left[\rho_j\right]\mathbb{E}\left[\rho_k\right] = Cov\left[\log\rho_j\log\rho_k\right] \end{aligned} \tag{3.41}$$

while also being equal to:

$$\partial_{\lambda_j}\partial_{\lambda_k}\log Z(\vec{\lambda})\Big|_{\vec{\lambda}=0} = \psi^{(1)}(N + a|\mathcal{K}|) - \delta_{jk}\psi^{(1)}(\hat{K}_j + a) \tag{3.42}$$

so that:

$$Cov[g^{\mu}g^{\nu}] = \frac{1}{2^{2n}}\sum_{j,k}\chi_j^{\mu}\chi_k^{\nu}(\psi^{(1)}(N + a|\mathcal{K}|) - \delta_{jk}\psi^{(1)}(\hat{K}_j + a))$$

We can further simplify by using a useful property of the χ matrix:

$$\sum_j \chi_j^{\mu} = \sum_j\sum_{s\in Q_j}\phi^{\mu}(s) = \sum_s^{\mathcal{S}}\phi^{\mu}(s) = 0 \tag{3.43}$$

(χ by definition does not contain the "$\mu = 0$" row!).
Thanks to this we can finally write:

$$Cov[g^{\mu}, g^{\nu}] = \frac{1}{2^{2n}}\sum_j \chi_j^{\mu}\chi_j^{\nu}\psi^{(1)}(\hat{K}_j + a) \tag{3.44}$$

If we now look at the *variance* of the "transformed" couplings $\tilde{g}^{\eta}$:

$$\begin{aligned} Var[\tilde{g}^{\eta}] =& \sum_{\mu}\sum_{\nu}U_{\mu\eta}U_{\nu\eta}\mathbb{E}\left[g^{\mu}g^{\nu}\right] - \left(\sum_{\mu}U_{\mu\eta}\mathbb{E}\left[g^{\eta}\right]\right) \\ =& \sum_{\mu}\sum_{\nu}U_{\mu\eta}U_{\nu\eta}Cov[g^{\mu}, g^{\nu}] \\ =& \frac{1}{2^{2n}}\sum_{\mu}\sum_{\nu}U_{\mu\eta}U_{\nu\eta}\sum_j\chi_j^{\mu}\chi_j^{\nu}\psi^{(1)}(\hat{K}_j + a) \\ =& \frac{1}{2^{2n}}\sum_{j,a,b}\Lambda_a\Lambda_b W_{aj}W_{bj}\sum_{\mu}U_{\mu\eta}U_{\mu a}\sum_{\nu}U_{\nu\eta}U_{\nu b}\psi^{(1)}(\hat{K}_j + a) \\ =& \left(\frac{\Lambda_{\eta}}{2^n}\right)^2\sum_j W_{\eta j}^2\psi^{(1)}(\hat{K}_j + a) \end{aligned} \tag{3.45}$$

We see that this is proportional to Λ_η, meaning that the average statistical error in the estimation of couplings g^η relative to our different sufficient statistics ϕ^η is proportional to the corresponding singular value of the χ matrix.
It would be tempting now, as pointed out in [6], to regard the statistics corresponding with largest Λ^η as so-called *sloppy modes*, in the sense of [15]; yet from our discussion in Chapter 1 we should be convinced that high parameter sensitivity to sampling noise is a symptom of overfitting - with respect to which we should be relatively safe, thanks to us having performed model selection in the first place. Resolution of this apparent paradox amounts to noting that *the values of g^η are proportional to Λ^η themselves* (cfr. 3.29). This means that across our family of sufficient statistics, *relative fluctuations* in the inferred parameters are of comparable magnitude.

Estimating $\vec{\rho}$ before insertion in 3.1

Instead of using the random vector $\vec{\rho}$, we can compute a posterior estimate $P(\mathbf{s}|\hat{s}_N, \mathcal{K})$ and insert this in 3.1 as our s-probability:

$$\begin{aligned}
P(\mathbf{s}|\hat{s}_N, \mathcal{K}) =& \int_{\vec{\rho}_\mathcal{K}} d\vec{\rho}_\mathcal{K} P(\mathbf{s}|\hat{s}_N, \vec{\rho}_\mathcal{K}, \mathcal{K}) P(\vec{\rho}_\mathcal{K}|\hat{s}_N, \mathcal{K}) \\
=& \int_{\vec{\rho}_\mathcal{K}} d\vec{\rho}_\mathcal{K} P(\mathbf{s}|\vec{\rho}_\mathcal{K}, \mathcal{K}) \frac{P(\hat{s}_N|\vec{\rho}_\mathcal{K}, \mathcal{K}) P(\vec{\rho}_\mathcal{K}|\mathcal{K})}{P(\hat{s}_N)} \\
=& \frac{1}{P(\hat{s}_N)} \int_{\vec{\rho}_\mathcal{K}} d\vec{\rho}_\mathcal{K} \frac{\rho_{j(s)}}{m_{j(s)}} \prod_r^\mathcal{S} \left(\frac{\rho_{j(r)}}{m_{j(r)}} \right)^{\hat{k}_r} P(\vec{\rho}_\mathcal{K}|\mathcal{K}) \\
\propto& \int_{\vec{\rho}_\mathcal{K}} d\vec{\rho}_\mathcal{K} \prod_j^\mathcal{K} \left(\frac{\rho_j}{m_j} \right)^{\hat{K}_j + \delta_{j,j(s)}} \cdot \frac{\Gamma(|\mathcal{K}|a)}{\Gamma(a)^{|\mathcal{K}|}} \prod_j^\mathcal{K} \rho_j^{a-1} \delta_{(\sum_j \rho_j - 1)} \\
=& \left[\frac{\Gamma(a|\mathcal{K}|)}{\Gamma(a)^{|\mathcal{K}|} \Gamma(N + a|\mathcal{K}|)} \frac{\prod_j \Gamma(\hat{K}_j + a)}{m_j^{\hat{K}_j}} \right] \frac{1}{m_{j(s)}} \frac{\hat{K}_{j(s)} + a}{N + a|\mathcal{K}|}
\end{aligned}$$

One can check (see chapter 5 for details) that normalization in this case amounts exactly to division by the quantity in parentheses. This way we end up with the posterior $\mathcal{K}$-estimate:

$$P(\mathbf{s}|\hat{s}_N, \mathcal{K}) = \frac{1}{m_{j(s)}} \frac{\hat{K}_{j(s)} + a}{N + a|\mathcal{K}|} \tag{3.46}$$

The reader familiar with classic Bayesian statistics will recognize in this expression an instance of *Laplace's rule of succession*. We now need to fix the prior parameter:

- for $a \to 0$ we get:

$$P(\mathbf{s}|\hat{s}_N, \mathcal{K}, \mathcal{I}_0) = \frac{1}{m_{j(s)}} \frac{\hat{K}_{j(s)}}{N} = \frac{\hat{k}_s}{N} \tag{3.47}$$

  which turns out to coincide with the maximum likelihood estimator; this is the prior used throughout all of [6].

- for $a = 1$ we get:

$$P(\mathbf{s}|\hat{s}_N, \mathcal{K}, \mathcal{I}_1) = \frac{1}{m_{j(s)}} \frac{\hat{K}_{j(s)} + 1}{N + |\mathcal{K}|} \tag{3.48}$$

  Which we will call a 1-*pseudocount posterior*: it amounts in fact to an empirical frequency estimation, where said frequency is corrected by attributing one fake observation to each "grouped" state.
  Pseudocount estimators are very popular in inference problems (and have in some cases be found to be a necessary choice; see [29]), because they tipically allow to overcome the problem of *divergencies caused by unobserved states*. In fact, if we plug our $a \to 0$ estimate into 3.1 we see immediately that for any state s_0 such that $\hat{k}_{s_0} = 0$ we have a "log(0)" divergence. Next section will be devoted to discussion of this issue.

3.6 Unobserved states and the need for regularization

Consider again our core formula:

$$g^\mu = \frac{1}{2^n} \sum_j \chi_j^\mu \log \frac{\rho_j}{m_j}$$

If we resort to maximum likelihood estimation for $\vec{\rho}_{\mathcal{Q}}$, this becomes:

$$g^\mu = \frac{1}{2^n} \sum_j \chi_j^\mu \log \frac{\hat{K}_j}{m_j N} = \frac{1}{2^n} \sum_j \chi_j^\mu \log \frac{\hat{k}_{s_j}}{N} \tag{3.49}$$

where s_j is a representative state belonging to the jth partition set: $s \in Q_j$.
Now we see that if there are states that are never observed ($\hat{k} = 0$) we get divergencies in the estimated values of g^μ - to be precise, we'll have divergent couplings for every μ such that $\chi_\varnothing^\mu \neq 0$.
The bad news is that in practice, we are more or less *always* in this situation, since a good sampling of a 2^n-sized alphabet would require in all interesting application huge sample sets, that we can't afford. Plus, these divergencies are in general mantained in the $\tilde{g}^\eta$-basis.

How can we then *regularize* these divergencies? Here we state two main possible recipes for doing so:

- The most popular regularizing devices are L1 (Lasso) and L2 (ridge) regression methods; they both amount to adding a penalty term to the cost function to be optimized (in our case - the loglikelihood of the sample) such that greater magnitudes of the couplings are penalized. L1 regression optimizes:

$$\mathcal{L}(\mathbf{g}) = \log P(\hat{s}_N|\mathbf{g}) - \alpha \sum_{\mu} |g^{\mu}| \tag{3.50}$$

  While L2 regression optimizes:

$$\mathcal{L}(\mathbf{g}) = \log P(\hat{s}_N|\mathbf{g}) - \alpha \sum_{\mu} (g^{\mu})^2 \tag{3.51}$$

- A method to avoid said divergencies would be that of *not* resorting to plain maximum likelihood estimation, thus mantaining partial bayesianity and for instance using the 1-pseudocount posterior estimates:

$$P(\mathbf{s}|\hat{s}_N, \mathcal{K}, \mathcal{I}_1) = \frac{1}{m_{j(s)}} \frac{\hat{K}_{j(s)} + 1}{N + |\mathcal{K}|} \tag{3.52}$$

  as described in the previous section. Pseudocounts ensure that no divercency whatsoever is possible. This is the most sound choice from a Bayesian point of view, and in turn a Bayesian point of view is arguably mandatory when we work in the deep undersampling regime, which is the one in which prior informations have the widest influence on our inference procedures.

3.7 Numerical results

We replicated some numerical results found in [6] and [5]. There, the authors applied the method discussed to the data used in [30] referring to the decision of a U.S. Supreme Court on 895 cases. This court is composed of nine judges, each expressing a binary vote with respect to a given case. The court is thus modeled as a $n = 9$ spin system, with a sample of 895 observations (cases). In [30] a graphical model was reconstructed while admitting only two-spin interactions; the aim in [6] was to understand whether the system at hand could be more "simply" described by admitting higher order interactions. Here, following [6], we perform inference both on this dataset and, for comparison, on a synthetic one generated from a fully connected pairwise $n = 9$ fully degenerate spin model, with a temperature chosen as to match the average two-spin correlations in the real dataset.

We provide estimated couplings $\mathbb{E}[\vec{g}]$ in different cases:

- Absence of regularization schemes - resulting in very high ($\rightarrow$ divergent) values of the inferred parameters;
- Regularization via pseudocounts (a=1)

Details on the dataset can be found in [30].

3.7.1 Synthetic data: full pairwise model

The chosen Hamiltonian is the one of a ferromagnetic Ising model:

$$\mathcal{H}_{\text{full pair}} = \sum_{i=1}^{9} \sum_{j<i} s_i s_j \tag{3.53}$$

(we use the sign convention: $P(s|g) = e^{+\frac{\beta}{n} \sum_\mu g^\mu \phi^\mu(s)}$); we set $\beta = 2.2$.

Sample frequencies

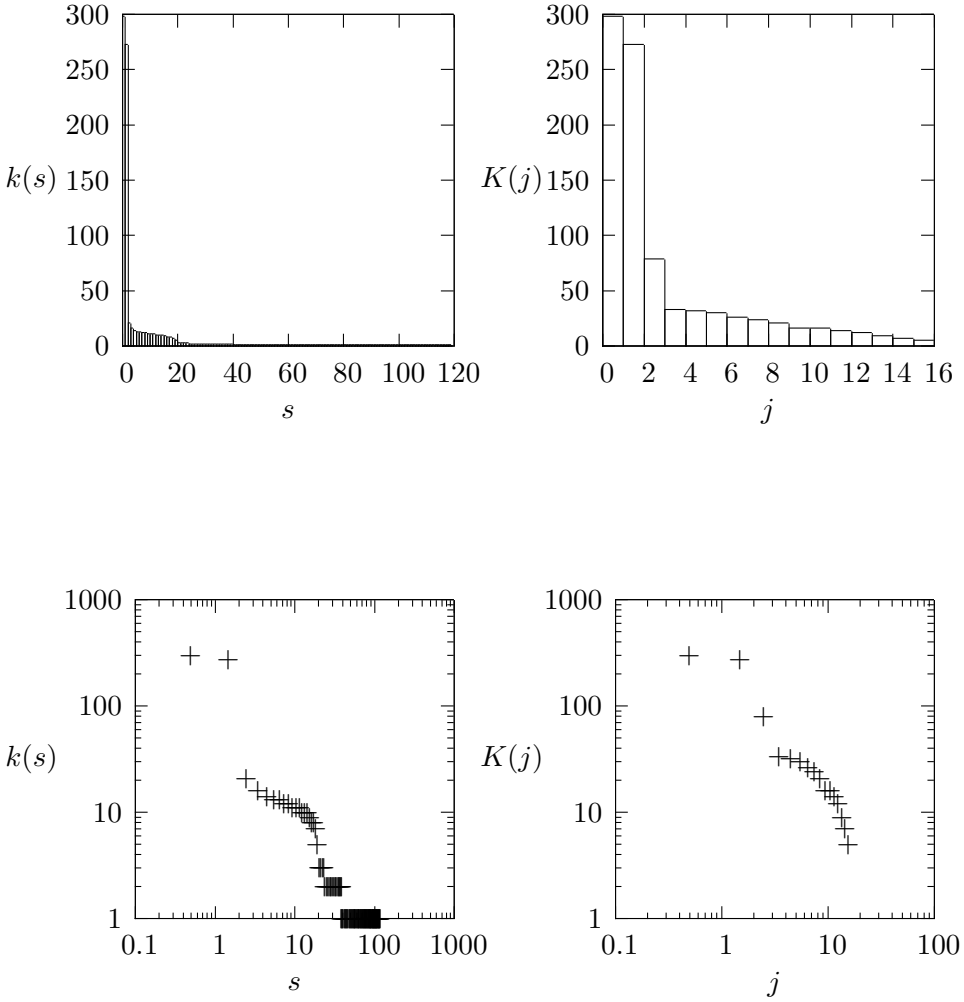

Figure 3.4: *Sample frequencies statistics.*
Top left: empirical frequencies of states plotted against their rank; notice how, in our dataset, we have only 120 observed states, out of $2^9 = 512$ possible ones. Top right: empirical frequencies $K(j)$ relative to the frquency partition set, as defined in Chapter 3, plotted against their rank. There are 16 different frequencies occurring in the sample, so that the $\mathcal{K}$ partition is composed of 16 sets. Bottom: same two plots in logarithmic scale.

Inferred couplings: orders of interactions

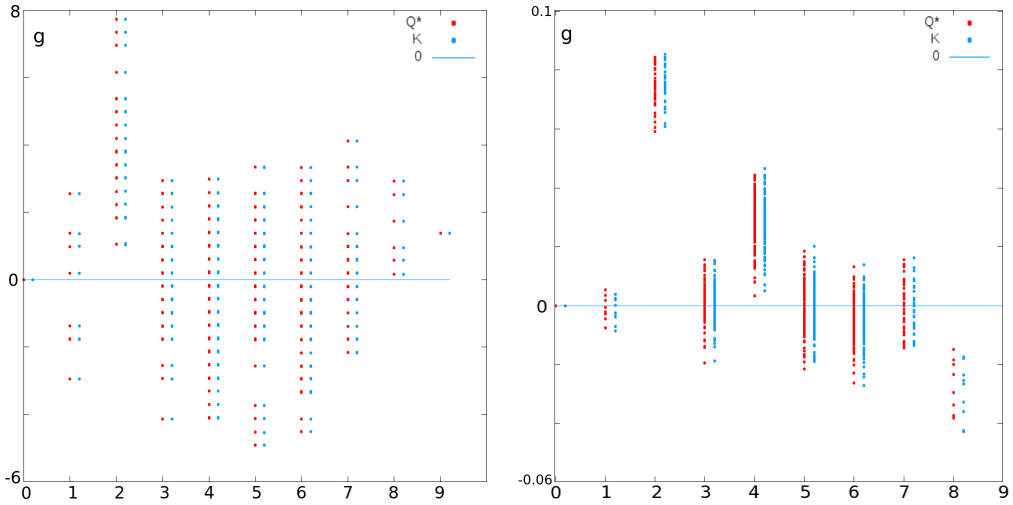

Figure 3.5: *Orders of inferred interactions*
Plots of the values of $\mathbb{E}[g^{\mu}]$ by their order as spin interactions. Left: No regularization performed; the values appear finite just because of the limits of numerical approximation. Right: "pseudocount"-regularized (a=1) case. Blue dots refer to parameters obtained under the $\mathcal{K}$ partition, red ones refer to the $\mathcal{Q}^*$ optimal one.

As we see in figure 3.5, non-regularized "$a = 0$" inference is extremely noisy; two-order interactions are prominent as desired, yet we have no sharp separation between these and the remaining ones. In the case of a pseudocount-regularized scheme, we see that the situation is far better.
We also see how while interactions of odd order are 0 on average, "spurious" interactions of even order emerge.
A last observation is that inference seems to be very lightly affected by passing to the optimal $\mathcal{Q}^*$ partition, at least in this low dimensional case. This reassures us about the fact that we can safely use, in practice, the frequency one.

The emergence of spurious interactions like the fourth-order ones in figure 3.5 is interesting because it is arguably robust against L1, L2 schemes for sparsification. Extensive numerical investigations on synthetic datasets generated using various models show that this phenomenon is general and order - independent. What happens is that *simple XOR combinations of the "true" operators tend to get high rankings*. When this happens, the obtained strength for a spurious interaction is proportional (in absolute value) to the strengths of the "true" inferred interactions that generated it via XOR combination.
A precise characterization of this effect is lacking, and will be the object of future investigations.

3.7.2 Real data: the U.S. Supreme court

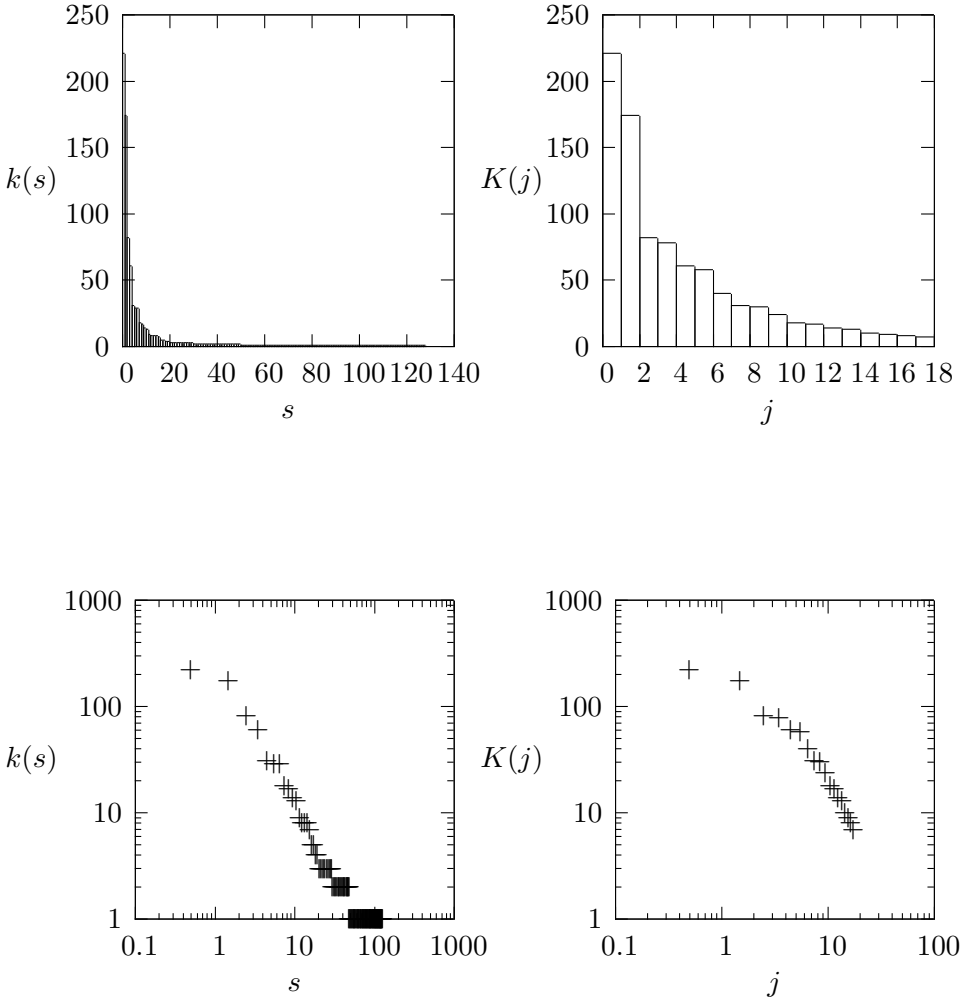

Figure 3.6: *Sample frequencies statistics.*
Top left: empirical frequencies of states plotted against their rank.
Top right: empirical frequencies $K(j)$ relative to the frquency partition set plotted against their rank.
Bottom: same two plots in logarithmic scale.

Inferred couplings: orders of interactions

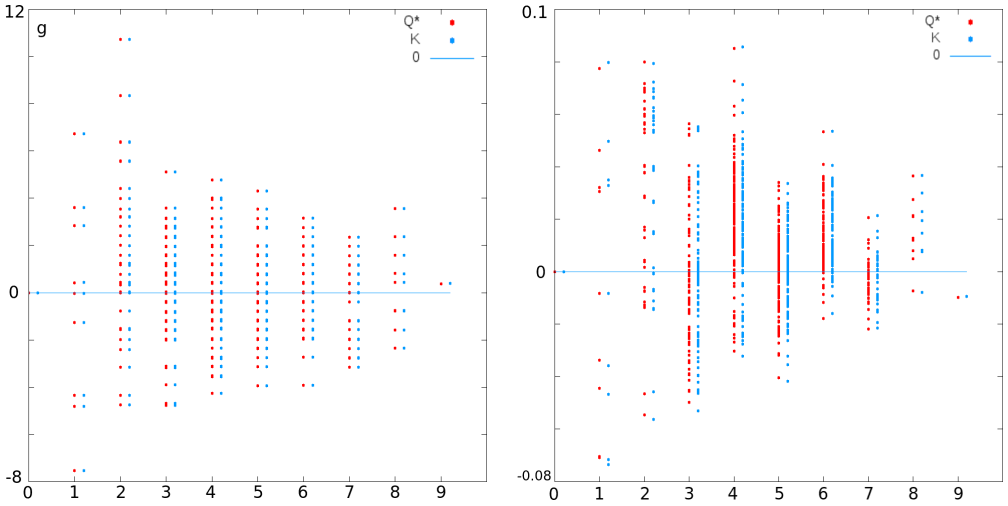

Figure 3.7: *Orders of inferred interactions*
Plots of the values of $\mathbb{E}[g^\mu]$ by their order as spin interactions. Top: No regularization performed; the values appear finite just because of the limits of numerical approximation. Bottom: "pseudocount"-regularized (a=1) case. Blue dots refer to parameters obtained under the $\mathcal{K}$ partition, red ones refer to the $\mathcal{Q}^*$ optimal one.

We clearly see that first and second order interactions ore prominent. Yet, in agreement with [6], we see that many fourth order interactions are in the top 5% couplings. This prediction comes from an uninformative Bayesian procedure, can arguably advocate for specific *relevance* of these interactions for the system.

Chapter 4

Mining the χ matrix

In this chapter, we will complete the exposition contained in the previous one, with the presentation of original findings about how the information about the selected partition $\mathcal{Q}$ is represented in the SVD decomposition of the χ matrix.
We will discover a nice correspondence between the singular values Λ_η and the sizes $|Q_j|$ of the partition sets, allowing us to argue that the choice of discarding the smallest singular values is actually *not* justified - in fact, this new interpretation suggests that a possible mean of regularization would be that of discarding the single one *biggest* singular value.

4.1 Analytical results for the χ matrix

We will devote this section to raw calculation, and pospone discussion of the results to the next one. The main results of next subsection are formulas 4.3 and 4.14. Details of their derivation can be skipped at a first reading.

4.1.1 $\{\psi^\eta\}_\eta$ as functions of (W,Λ,$\mathcal{Q}$)

We can write some meaningful identities starting from the definition of the sufficient statistics ψ^η.
First, we compute, $\forall \eta : \Lambda_\eta \neq 0$:

$$\begin{aligned}
\psi^\eta(s) &= \sum_{\mu>0} \phi^\mu(s) U_{\mu\eta} = \sum_{\mu>0} \phi^\mu(s) \sum_j \chi_{\mu j} W_{\eta j} \Lambda_\eta^{-1} \\
&= \sum_j W_{\eta j} \Lambda_\eta^{-1} \sum_{\mu>0} \phi^\mu(s) \sum_{r \in Q_j} \phi^\mu(r) \\
&= \sum_j W_{\eta j} \Lambda_\eta^{-1} \sum_{r \in Q_j} \left(\sum_{\mu \geq 0} \phi^\mu(s) \phi^\mu(r) - 1 \right)
\end{aligned}$$

Due to the orthogonality of the spin operators:

$$\begin{aligned}\psi^{\eta}(s) &= \sum_j W_{\eta j}\Lambda_{\eta}^{-1}\left[\sum_{r\in Q_j}(\delta_{r,s}2^n) - |Q_j|\right] \\ &= \sum_j W_{\eta j}\Lambda_{\eta}^{-1}(\delta_{j,j(s)}2^n - |Q_j|) \\ &= \frac{2^n W_{\eta j(s)}}{\Lambda_{\eta}} - \frac{\sum_j W_{\eta j}|Q_j|}{\Lambda_{\eta}}\end{aligned} \tag{4.1}$$

that, if we define:

$$\omega_j \equiv \frac{|Q_j|}{2^n}, \quad \lambda_{\eta} = \left(\frac{\Lambda_{\eta}}{2^n}\right)^2, \quad \overline{W_{\eta}} \equiv \sum_j \omega_j W_{\eta j}$$

becomes:

$$\psi^{\eta}(s) = \frac{W_{\eta j(s)} - \overline{W_{\eta}}}{\sqrt{\lambda_{\eta}}} \qquad \forall \eta : \Lambda_{\eta} \neq 0 \tag{4.2}$$

This is an important result! It demonstrates how *the sufficient statistics ψ^{η}, as functions of the alphabet, only depend on the states through their frequencies*:

$$\psi^{\eta}(s) = \psi^{\eta}(j(s)) \tag{4.3}$$

"j" denoting here a specific partition set. This key observation enables various further characterizations; as a first thing, we can say that this equation must also be true:

$$\begin{aligned}\psi^{\eta}(s) = \psi^{\eta}(j(s)) &= \frac{1}{|Q_{j(s)}|}\sum_{r\in Q_{j(s)}}[\psi^{\eta}(r)] = \\ &= \frac{1}{|Q_{j(s)}|}\sum_{r\in Q_{j(s)}}\sum_{\mu>0}\phi^{\mu}(r)U_{\mu\eta} = \\ &= \frac{1}{|Q_{j(s)}|}\sum_{\mu>0}U_{\mu\eta}\chi_{\mu j(s)} = \frac{\sqrt{\lambda_{\eta}}W_{\eta j(s)}}{\omega_{j(s)}}\end{aligned}$$

Which we'll write, forgetting s, as:

$$\psi^{\eta}(j) = \frac{\sqrt{\lambda_{\eta}}W_{\eta j}}{\omega_j} \tag{4.4}$$

4.1.2 $\{W_{\eta j}\}_{\eta,j}$ as functions of ($\{\lambda_\eta\}_\eta$, $\{\omega_j\}_j$)

Putting together (4.2) and (4.4) we see that, $\forall \eta : \lambda_\eta \neq 0$, $\forall j$:

$$\frac{W_{\eta j} - \overline{W_\eta}}{\sqrt{\lambda_\eta}} \equiv \frac{\sqrt{\lambda_\eta} W_{\eta j}}{\omega_j}$$

from which:

$$\lambda_\eta W_{\eta j} = \omega_j \left(W_{\eta j} - \overline{W_\eta} \right) \tag{4.5}$$

Summing over j:

$$\lambda_\eta \sum_j W_{\eta j} = \overline{W_\eta} - \overline{W_\eta} = 0$$

meaning that, since $\lambda_\eta \neq 0$ by hypothesis:

$$\sum_j W_{\eta j} = 0 \quad \forall \eta : \Lambda_\eta \neq 0 \tag{4.6}$$

Also, from (4.5) we get:

$$W_{\eta j} \left[\frac{\omega_j - \lambda_\eta}{\omega_j} \right] = \overline{W_\eta} \tag{4.7}$$

from which we see that the expression at left hand side must be constant with respect to j.
We also see that whenever we choose η such that λ_η coincides with ω_j for some j, we get:

$$\overline{W_\eta} = 0 \qquad \forall \eta : \exists j : \omega_j \equiv \lambda_\eta \tag{4.8}$$

Whereas for other choices of η we can write:

$$W_{\eta j} = \frac{\omega_j \overline{W_\eta}}{\omega_j - \lambda_\eta} \qquad \forall \eta : \nexists j : \omega_j \equiv \lambda_\eta \tag{4.9}$$

Since W is unitary, we can impose normalization of its rows:

$$1 = \sum_j W_{\eta j}^2 = \overline{W_\eta}^2 \sum_j \left(\frac{\omega_j}{\omega_j - \lambda_\eta} \right)^2 \qquad \forall \eta : \nexists j : \omega_j \equiv \lambda_\eta$$

meaning:

$$\overline{W_\eta} = \frac{1}{\sqrt{\sum_k \left(\frac{\omega_k}{\omega_k - \lambda_\eta} \right)^2}} \qquad \forall \eta : \nexists j : \omega_j \equiv \lambda_\eta \tag{4.10}$$

so that (4.8) and (4.10) give us $\overline{W_\eta}$ in all cases.
Substituting (4.10) in (4.9) we get all "non degenerate" W elements as functions of the spectrum, and of the partition sets' cardinalities:

$$W_{\eta j} = \frac{\frac{\omega_j}{\omega_j - \lambda_\eta}}{\sqrt{\sum_k \left(\frac{\omega_k}{\omega_k - \lambda_\eta}\right)^2}} \qquad \forall \eta : \nexists j : \omega_j \equiv \lambda_\eta \tag{4.11}$$

Giving us the "non-degenerate" sufficient statistics, as a function of χ's spectrum and $\mathcal{Q}$-sets' cardinalities only:

$$\psi^\eta(j) = \frac{\frac{\lambda_\eta}{\omega_j - \lambda_\eta}}{\sqrt{\sum_k \left(\frac{\omega_k}{\omega_k - \lambda_\eta}\right)^2}} \qquad \forall \eta : \nexists j : \omega_j \equiv \lambda_\eta \tag{4.12}$$

Note that, $\forall \eta : \nexists k : \lambda_\eta = \omega_k$, the entries $W_{\eta j}$ (and so $\psi^\eta(j)$ as well), are constant along "degenerate" components:

$$W_{\eta j} = c_\eta(\omega) \quad \forall j \in \mathcal{J}_\omega \equiv \{j \mid \omega_j = \omega\} \tag{4.13}$$

4.1.3 The spectrum $\{\lambda_\eta\}_\eta$

The main contribution to our comprehension of the spectrum comes from (4.9): multiplying both sides by ω_j and summing over j:

$$1 = \sum_j \frac{\omega_j^2}{\omega_j - \lambda_\eta} \qquad \forall \eta : \nexists j : \omega_j \equiv \lambda_\eta \tag{4.14}$$

This last formula allows us to investigate and characterize properly the spectrum of singular values; for a given choice of $\vec{\omega}$ on the $|\mathcal{Q}|$-dimensional probability simplex, the shape of the right hand side, seen as a function of λ, is depicted in figure 4.1; there, divercengies correspond to values of λ such that $\lambda \equiv \omega_j$ for some j.
If we order the indices η by the increasing value of Λ_η (and thus λ_η), and the indices j by the increasing value of the cardinalities $|Q_j|$ (and thus ω_j), we can immediately write down the properties:

$$\begin{aligned} &\lambda_0 \equiv 0 \\ &\lambda_k \in [\omega_k, \omega_{k+1}] \qquad \forall k : \nexists j : \omega_j \equiv \lambda_k \end{aligned} \tag{4.15}$$

or, equivalently:

$$\begin{aligned} &\Lambda_0 = 0 \\ &\Lambda_k \in \left[\sqrt{2^n |Q_k|}, \sqrt{2^n |Q_{k+1}|}\right] \quad \forall k : \nexists j : |Q_j| \equiv \frac{\Lambda_k^2}{2^n} \end{aligned} \tag{4.16}$$

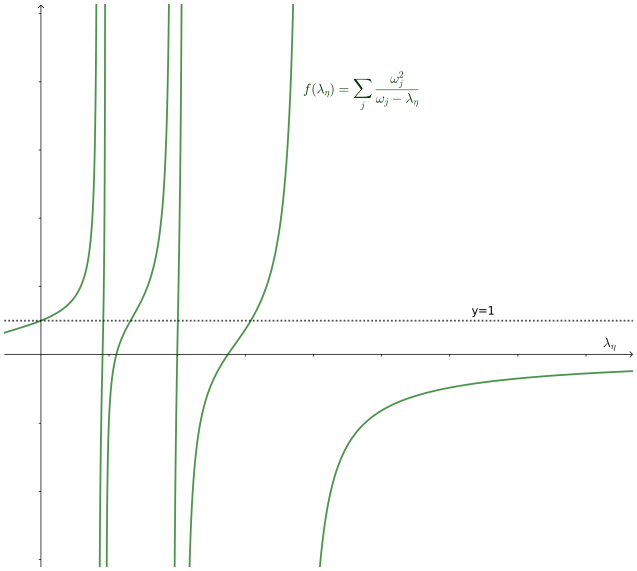

Figure 4.1: Plot of 4.14
the values λ_η correspond to intersections between the green function and the dashed line $y = 1$. One has vertical asymptotes for values $\lambda = \omega_j$, and one λ_η solution between each pair of these lines; if $\omega_j = \omega_k$ for some j, k, the corresponding solution is $\lambda_\eta = \omega_j = \omega_k$.

Exhamine the plot. If all the cardinalities $|Q_j|$ are different we get all the $|\mathcal{Q}| - 1$ nonzero singular values trapped in $|\mathcal{Q}| - 1$ of the above intervals, one for each singular value, whereas everytime we make the cardinality of two different partition sets coincide the corresponding interval (always containing one of the singular values) gets "squished" inside a single divergency. By this line of reasoning we conclude that everytime we have degeneracies in the cardinalities we'll observe degeneracy in singular values: more precisely, we'll observe $|Q_{j_{deg}}| - 1$ coincident singular values, all of value $\lambda_{j_{deg}} = \omega_{j_{deg}}$. Further characterization of the spectrum can be found in Appendix C.

4.1.4 The "degenerate" rows of the W matrix

Consider (4.7) for some j (if it exists) for which $\omega_j \neq \lambda_\eta$: the expression in square brackets is, in this case, $\neq 0$; yet we know from (4.8) that $\overline{W_\eta} = 0$. This implies:

$$W_{\eta j} \equiv 0 \qquad \forall(\eta, j) : (\exists k : \omega_k \equiv \lambda_\eta) \wedge (\omega_j \neq \lambda_\eta) \tag{4.17}$$

. In turn, we cannot determine the values of W for all other values of j. For each "degenerate" $|Q_k|$ we'll thus have $|Q_k - 1|$ "localized" (they have nonzero components only along the corresponding degenerate partition sets), singular vectors, forming an orthonormal set (due to the above properties);

this set can be completed to become a basis of the linear space spanned unconstrainedly by those nonzero component, by adding a vector which is constant along those directions, and zero elsewhere.
Indeed, the $\mathbf{\Lambda_0 = 0}$ row is a constant row - since W is unitary, and its other rows span the linear subspace orthogonal to the $(1, 1, \cdots, 1)$ vector; its entries are :

$$W_{0j} = \frac{1}{\sqrt{|\mathcal{Q}|}} \tag{4.18}$$

4.1.5 W-redefinition of the sufficient statistics

Let's see what happens when we write the maximum entropy distribution for the ψ^η and enforce normalization:

$$P(s|g) = \frac{1}{Z_g} e^{\sum_\eta g^\eta \psi^\eta(s)} \tag{4.19}$$

Now, using (4.1):

$$\begin{aligned} P(s|g) &= \frac{1}{Z_g} e^{-\sum_\eta \frac{g^\eta}{\Lambda_\eta} 2^n \overline{W_{\eta j}}} e^{\sum_\eta \frac{g^\eta}{\Lambda_\eta} 2^n W_{\eta j(s)}} \\ &= \frac{1}{Z_c} e^{\sum_\eta c^\eta \zeta^\eta(j(s))} \end{aligned} \tag{4.20}$$

where

$$\begin{aligned} \zeta^\eta(j) &\equiv \sqrt{|\mathcal{Q}|} W_{\eta j} \\ c^\eta &\equiv \frac{g^\eta}{\sqrt{\lambda_\eta |\mathcal{Q}|}} \end{aligned} \tag{4.21}$$

This equivalent redefinition of our sufficient statistics is nice because, since W is unitary, we now see that the ζ^η functions obey the same orthogonality and completeness relations as the old basis functions ϕ^μ:

$$\begin{aligned} \sum_\eta \zeta^\eta(j)\zeta^\eta(k) &= |\mathcal{Q}|\delta_{j,k} \\ \sum_j \zeta^\eta(j)\zeta^\rho(j) &= |\mathcal{Q}|\delta_{\eta,\rho} \end{aligned} \tag{4.22}$$

Also, we can check via 3.45 that the variance of the inferred parameters in this representation is no more modulated by the singular values:

$$\begin{aligned} Var[c^\eta] =& \frac{1}{\lambda^\eta |\mathcal{Q}|} Var[g^\eta] = \frac{1}{\lambda^\eta |\mathcal{Q}|} \lambda^\eta \sum_j W_{\eta j}^2 \psi^{(1)}(\hat{K}_j + a) \\ =& \frac{1}{|\mathcal{Q}|} \sum_j W_{\eta j}^2 \psi^{(1)}(\hat{K}_j + a) \end{aligned} \tag{4.23}$$

If we fix a prior parameter $a > 1.5$, so that $\psi^{(1)}(\hat{K}_j + a)$ is always positive, it is meaningful to write the bound:

$$Var[c^\eta] < \frac{\psi^{(1)}(\hat{K}_j^{\max} + a)}{|\mathcal{Q}|} \tag{4.24}$$

4.2 Pause and ponder: a new method for regularization?

Let's pause now and try to understand what significance there is in last section's findings: from our point of view, the main achievement is the discovery of relation (4.14) for the magnitudes of χ's singular values λ_η. In particular, this relation tells us that the *biggest* singular value $\lambda^{\max}$ correlates with the partition set containing the highest number of states:

$$Q_{j_{\max}}, \qquad\qquad j_{\max} = \arg\max_j m_j$$

The typical statistical features of samples in the undersampling regime are such that the most populated partition set usually coincides with the set of unobserved states.
This hints that maybe an efficient strategy for regularization is that of discarding the *biggest* singular value, in place of the smaller ones (and thus "stop listening" to unobserved states). Exploration of this path is a perspective for future research.

4.3 Bonus: reverse engineering partitions & hidden loop structures

It is clear that the interface between mixtures and spin models hides plenty of structure; here, we want to highlight a nice result emerging from some experimenting with mappings between these two classes.
What follows has not yet been properly formalized; we thus adopt a more informal style, giving first a couple examples to gain some insight, then presenting a somewhat general empirical observation. This all must be intended as a hint towards future investigation.

4.3.1 Spin to mixture constructive recipe: examples

The following is motivated by a simple question. We have seen how to map mixture models into spin models, via the χ matrix associated to a partition $\mathcal{Q}$. Now pretend we don't know any of this for a bit and ask ourselves: given a *desired* partitioning of a set of states, how can we construct a spin model

that returns probabilities of states generating precisely that partitioning? Let's start from scratch. The empty Hamiltonian:

$$\mathcal{H}_0(s) = 0 \tag{4.25}$$

generates a uniform distribution $P(s) = \frac{1}{|\mathcal{S}|}$. Thus, the corresponding partition is the "single-chunk" one $\mathcal{Q}_0$. If we add $(+g^\mu \phi^\mu(s))$ to our Hamiltonian:

$$\mathcal{H}_0(s) = g^\mu \phi^\mu(s) \tag{4.26}$$

we induce a partitioning of states in two equally sized classes:

- the set $\{s^+\}$ of states for which $\phi^\mu(s^+) = +1$;
- the set $\{s^-\}$ of states for which $\phi^\mu(s^-) = -1$;

where $\{s^+\} \bigcup \{s^-\} \equiv \mathcal{S}$.
Let's see, for example, what happens for a 4-spin model:

s	$\mathcal{H}_0(s) = 0$
$++++$	
$+++-$	
$++-+$	
$++--$	
$+-++$	
$+-+-$	
$+--+$	
$+---$	$\mathcal{H} = 0$
$-+++$	
$-++-$	
$-+-+$	
$-+--$	
$--++$	
$--+-$	
$---+$	
$----$	

$\longrightarrow$

s	$\mathcal{H}(s) = g^1 \phi^1(s)$
$++++$	
$+++-$	
$++-+$	
$++--$	
$+-++$	$\mathcal{H} = g^1$
$+-+-$	
$+--+$	
$+---$	
$-+++$	
$-++-$	
$-+-+$	
$-+--$	
$--++$	$\mathcal{H} = -g^1$
$--+-$	
$---+$	
$----$	

If we now "turn on" another operator, say, $\phi^2(s)$:

s	$\mathcal{H}_0(s)=1$	$\mathcal{H}(s)=\phi^1(s)$	$\mathcal{H}(s)=\phi^1(s)+\phi^2(s)$
$++++$ $+++-$ $++-+$ $++--$	$\mathcal{H}=0$	$\mathcal{H}=g^1$	$\mathcal{H}=g^1+g^2$
$+-++$ $+-+-$ $+--+$ $+---$			$\mathcal{H}=g^1-g^2$
$-+++$ $-++-$ $-+-+$ $-+--$		$\mathcal{H}=-g^1$	$\mathcal{H}=-g^1+g^2$
$--++$ $--+-$ $---+$ $----$			$\mathcal{H}=-g^1-g^2$

we split each of the previous partition sets in half. Thus it seems that we have a general recipe to split state spaces in 2^k distinct sets. Moreover, if we now decide to impose some *degeneracy*, for instance the additional constraint $g^1=g^2=g$, we manage to *merge* partition sets together: in fact we see that our $(4,4,4,4)$ partition becomes a $(4,8,4)$ one, because the Hamiltonians of the middle two partition sets become both identically $=0$. Let's add one more operator! We'll make two different choices: the natural choice ϕ^3 and also the choice $\phi^{(1,2)}$. Let's check what happens:

s	$g^1\phi^1(s)+g^2\phi^2(s)$ $+g^3\phi^3(s)$
$++++$ $+++-$	$+g^1+g^2+g^3$
$++-+$ $++--$	$+g^1+g^2-g^3$
$+-++$ $+-+-$	$+g^1-g^2+g^3$
$+--+$ $+---$	$+g^1-g^2-g^3$
$-+++$ $-++-$	$-g^1+g^2+g^3$
$-+-+$ $-+--$	$-g^1+g^2-g^3$
$--++$ $--+-$	$-g^1-g^2+g^3$
$---+$ $----$	$-g^1-g^2-g^3$

s	$g^1\phi^1(s)+g^2\phi^2(s)$ $+g^{(12)}\phi^{(12)}(s)$
$++++$ $+++-$ $++-+$ $++--$	$+g^1+g^2+g^{(12)}$
$+-++$ $+-+-$ $+--+$ $+---$	$+g^1-g^2-g^{(12)}$
$-+++$ $-++-$ $-+-+$ $-+--$	$-g^1+g^2-g^{(12)}$
$--++$ $--+-$ $---+$ $----$	$-g^1-g^2+g^{(12)}$

We see that while in the first case we have a further splitting of the states, in the second case we are not modifying the partition. A little thinking convinces us of the basic fact that *adding an operator which is not* independent *from the ones already present in the Hamiltonian does not change the corresponding partition.* Here "independency" is to be intended in the sense specified in chapter 2.
We can play around with degeneracies in both cases. What one sees is that:

- Enforcing constraints in an *independent* Hamiltonian, as is the one on the left, can merge sets together, but cannot produce sets of cardinality different than a power of two;
- Particular constraints in a *dependent* Hamiltonian of the type on the right can produce sets with richer cardinalities.

To verify this second claim, let's require: $g^1 = g^2 = -g^{(12)} = g$. We get:

s	$g(\phi^1 + \phi^2 - \phi^{(12)})$
$++++$	$+g$
$+++-$	
$++-+$	
$++--$	
$+-++$	$-g$
$+-+-$	
$+--+$	
$+---$	
$-+++$	
$-++-$	
$-+-+$	
$-+--$	
$--++$	
$--+-$	
$---+$	
$----$	

We thus managed to create a partition set composed of 12 elements. This would have not been possible with just independent operators.

4.3.2 General construction

Let's try and be more general now. We gained some useful insight but we are very far from having gained the ability of fully manipulating partitions through specific choices of spin models.
We will proceed one step at a time. We have seen that "single-operator" Hamiltonians correspond to splittings of the set of states in two halves, each of cardinality 2^{n-1}. There are $\frac{1}{2}\binom{2^n}{2^{n-1}}$ possible such splittings (the order in

which we place partition sets is irrelevant), and these are many more than the $2^n - 1$ operators we can use. This means that we can only reproduce via single operator Hamiltonians a small portion of all possible half splittings - yet, *if we could actually recreate all such half splittings, we would be able to recreate any possible partition* just by easily predeterminable combinations of the half-splitting Hamiltonians. Let's see why this must be true:

- First, choose a state s. Let's say we want a half-splitting such that **s** ends up in the set of the two corresponding to the highest value of the Hamiltonian; there will be, in total, $\binom{2^n-1}{2^n}$ such splittings. We will say that these splittings *boost* the state s.
- Assume we have access to the Hamiltonians corresponding to splittings that boost s, each with couplings normalized such that half of the states have energy (+1) and the other half have energy (-1).
- Then we can *add* all the s-boosting Hamiltonians together to get a bigger one $\mathcal{H}_s$: this implies that state s gets boosted $\binom{2^n-1}{2^n}$ times, and acquires energy $\mathcal{H}_s(s) = \binom{2^n-1}{2^n}$, while all other states get by symmetry boosted the same number of times (precisely: $\mathcal{H}_s(r) = \frac{1}{2^n-1}\binom{2^n-1}{2^{n-1}}$).
- We have thus generated a partition in two sets, one of which is composed of state s alone. It is easy to see that if we can do this for one state, we can create any possible desired partitioning, just by suitable linear combinations of $\mathcal{H}_s$ hamiltonians for all possible states s.

We then are now convinced that half-splittings are enough for construction of any possible partition. The problem remains of how to find Hamiltonians $\mathcal{H}_{\text{half}}$ for half splittings.
We conducted by hand investigations in systems with a low number of spins, and found that, in all cases, *Hamiltonians $\mathcal{H}_{half}$ are composed by families of operators forming loops* (recall the definition of loop from Chapter 2).

Example

Let's see an example of this; consider a 3-spins system. There are $|\mathcal{S}| = 8$ states, and $\frac{1}{2}\binom{8}{4} = 35$ different partitions of the states space in half. The 7 possible single-operator models lead to 7 of these partitions: $(+++, ++-, +-+, +--), (\cdots) \to \{\phi^{(1)}\}$
$(+++, ++-, -++, -+-), (\cdots) \to \{\phi^{(2)}\}$
$(+++, ++-, --+, ---), (\cdots) \to \{\phi^{(12)}\}$
$(+++, +-+, -++, --+), (\cdots) \to \{\phi^{(3)}\}$
$(+++, +-+, -+-, ---), (\cdots) \to \{\phi^{(13)}\}$
$(+++, +--, -++, ---), (\cdots) \to \{\phi^{(23)}\}$
$(+++, +--, -+-, --+), (\cdots) \to \{\phi^{(123)}\}$

The remaining 28 partitions are obtained under degenerate "loop" models comprised of 1 spin operator ϕ^{μ} with corresponding coupling constant g^{μ}, and three spin operators ϕ^{ν_1},ϕ^{ν_3},ϕ^{ν_3} such that $\phi^{\nu_1 \oplus \nu_2 \oplus \nu_3} = \phi^{\mu}$, all with coupling constant $-g^{\mu}$.
$(+++, ++-, +--, ---), (\cdots) \rightarrow \{+\phi^{(3)}, -\phi^{(1)}, -\phi^{(2)}, -\phi^{(12)}\}$
There are 7 possible 4-loops; we get 28 different partitions because for each loop we have to choose which operator will have a sign opposite to the others, this giving us $7 \cdot 4$ possible choices. A global inversion of the signs of operators does not change the partition (it simply reverses the energies of different partition sets). Also, reversing the sign of only the odd operators is equivalent to reversing all the spin signs in my alphabet; this changes our partition in the one with exactly opposite states.

This finding, that all half partitionings are generated by loops, has been verified as true in a multiplicity of cases. Yet, we lack full characterization of the "loop" picture: this path will be investigated in future research. It is nice to see loops pop out from the mixture - spin model mapping: this, together with the fact that, as we have seen, we can always reexpress the spin partition function $\mathcal{Z}_{\mathcal{M}}(\mathbf{g})$ of a spin model in the "loop expansion" manner, hints for a possibility of devising a procedure of model selection in the "loop representation".

4.3.3 Relation with the χ matrix

Now, let's reconsider last section's results through the glass of Chapter 3's findings. We have seen that, in general, we can generate a desired partition by *boosting* separately its different component sets. We have also seen that the boosting Hamiltonians involved are composed of particular sums of loops.
Now, we must realize that all these findings must be represented, in some way, in the structure of the χ matrix, since this matrix constitutes itself the mapping from a desired partition to the corresponding spin model. In particular, we see from 3.1 that single partition sets are associated with single *columns* of the χ matrix. What we realize is this these columns *coincide* (up to rescalings) with the boosting Hamiltonians we could sistematically build in the manner just described.

We thus can try and look for loop-like structures inside the χ matrix itself. A possible path in this sense starts from the fact that, as we have seen, our sufficient statistics depend on the states only through the partition

sets they belong to:

$$\begin{aligned} j = j(s) \equiv j(r) &\Longrightarrow \psi^\eta(s) = \psi^\eta(r) \\ &\Longrightarrow \sum_{\mu>0} U_{\mu\eta}\phi^\mu(s) = \sum_{\mu>0} U_{\mu\eta}\phi^\mu(r) \\ &\Longrightarrow \sum_{\mu>0} U_{\mu\eta}(\phi^\mu(s) - \phi^\mu(r)) = 0 \end{aligned} \tag{4.27}$$

From here:

$$\begin{aligned} \sum_{s\in Q_j} (\phi^\nu(s)\cdot 0) = 0 = \sum_{s\in Q_j} \phi^\nu(s) \sum_{\mu>0} U_{\mu\eta}(\phi^\mu(s) - \phi^\mu(r)) = \\ = \sum_{\mu>0} U_{\mu\eta} \left[\sum_{s\in Q_j} \phi^\nu(s)\phi^\mu(s) - \phi^\mu(r) \sum_{s\in Q_j} \phi^\nu(s) \right] \\ = \sum_{\mu>0} U_{\mu\eta} \left[\sum_{s\in Q_j} \phi^{\nu\oplus\mu}(s) - \phi^\mu(r)\chi_j^\nu \right] \\ = \sum_{\mu>0} U_{\mu\eta} \left[\chi_j^{\nu\oplus\mu} - \phi^\mu(r)\chi_j^\nu \right] \end{aligned} \tag{4.28}$$

Then, summing over r:

$$\begin{aligned} 0 &= \sum_{\mu>0} U_{\mu\eta} \left[|Q_j|\chi_j^{\nu\oplus\mu} - \chi_j^\mu\chi_j^\nu \right] \\ &= \sum_{\mu>0} U_{\mu\eta} \left[\chi_j^0\chi_j^{\nu\oplus\mu} - \chi_j^\mu\chi_j^\nu \right] \\ &\equiv \sum_{\mu>0} U_{\mu\eta} A_{(j)}^{\mu\nu} \end{aligned} \tag{4.29}$$

What one sees numerically *in elementary cases* is that:
1) diagonalizing the matrix A_j leads to only $|Q_j| - 1$ nonzero eigenvalues, all of value $2^n|Q_j|$;
2) $A_{(j)}^{\mu\nu}$ is, in very simple cases, different from zero only for particular values of ν. This values form a loop (actually, at least two distinct loops l_j^+ and l_j^-, each composed of constant numerical entries).
A is symmetric, so property (2) implies that also, keeping fixed ν, the matrix is different from 0 only on values of μ in these two loops. This hints at a possibility of detailed characterization of the U matrix in terms of loops that would be really useful, for it may lighten the computational process of re-mapping of the ψ^η functions back on the spin representation. These relations will be investigated in future research.

Chapter 5

The $\mathcal{Q}$-expansion

Introduction

Reconsider for a second the various possibilities for the $p(s)$ estimate to be plugged into our core forumla 3.1.
Until now, we mainly worked with the *frequency partition*, thus using

$$P(s|\hat{s_N}, \mathcal{K}(\hat{s}_N), I_0) = \frac{k_s}{N} \tag{5.1}$$

as a posterior estimate; as we have seen, this can be interpreted both as a maximum likelihood estimate, and as a posterior average $\mathbb{E}_{P(\vec{\rho}|\hat{s}_N,\mathcal{K})}\left[\rho_{j(s)}\right]$ over the $|\mathcal{K}|$-probability simplex (each dimension of which represents a set of the frequency partition induced on the alphabet by the sample $\hat{s}_N$) when using a symmetric $a \to 0$ Dirichlet prior $P_{a\to 0}(\rho_{j(s)}|\mathcal{K})$.

We can also keep conditioning on $\mathcal{K}(\hat{s}_N)$ as being the "right" partition, but change the parameter in the prior, obtaining a "pseudocount" posterior estimate:

$$P(s|\hat{s_N}, \mathcal{K}(\hat{s}_N), I_a) = \frac{k_s + a}{N + a|\mathcal{K}|} \tag{5.2}$$

Another possible choice would be using $P(s|\hat{s_N}, \mathcal{Q}^*(\hat{s_N}), I_a)$, i.e. the posterior estimate obtained from a symmetric Dir-prior on the Q*-simplex - Q* being the "optimal" partition maximizing $P(\mathcal{Q}|\hat{s}_N)$.

In this chapter, we adopt a fully Bayesian point of view. This means that we don't fix any particular partition, and just expand the posterior estimate on the basis of all the possible partitions of the alphabet:

$$P(s|\hat{s_N}) = \sum_{\mathcal{Q}} P(s|\hat{s}_N, \mathcal{Q}) P(\mathcal{Q}|\hat{s}_N) \tag{5.3}$$

From here, we show that only the terms of this sum corresponding to partitions that are *not coarser* than the $\mathcal{K}$ one affect the selection of sufficent statistics. Numerical simulations in [25] suggest that these finer partitions have low posterior probabilities. We show analytically that in the regime in which this probabilities are *very* low we can model their effect on inference as generating an L_2 regularizator, thus providing a Bayesian justification for the introduction of such term.
We then run simulations to try and argue if this is the regime in which we work in practical applications.

5.1 The Q-expansion

Let's exhamine separately the two conditional probabilities appearing in 5.3: To keep things light, we defer details of computations to Appendix D. For the posterior probabilities of states we get (see D.0):

$$P(s|\hat{s}_N, \mathcal{Q}) = \frac{1}{m_{\mathcal{Q}j(s)}} \frac{K_{\mathcal{Q}j(s)} + a_{\mathcal{Q}}}{N + a_{\mathcal{Q}}\mathcal{Q}} \tag{5.4}$$

which is an instance of laplace's rule of succession for model $\mathcal{Q}$.

Model posteriors $P(\mathcal{Q}|\hat{s}_N)$

As seen in Chapter 3, for $a_{\mathcal{Q}} = a = 1$ (see D.1 for discussion of the case $a \to 0$):

$$P(\mathcal{Q}|\hat{s_N}) = \frac{P_0(\mathcal{Q})}{P_0(\hat{s_N})} \frac{(Q-1)!}{(N+Q-1)!} \prod_{j=1}^{Q} \left[\frac{K_j!}{m_j^{K_j}} \right] \tag{5.5}$$

If we now take a set Q_k of $\mathcal{Q}$ and "split" it in two separate sets Q_{k1}, Q_{k2}, we obtain a model $\mathcal{Q}_{\text{split}}$ whose posterior probability we can compare with the one of $\mathcal{Q}$:

$$\frac{P(\mathcal{Q}_{split}|\hat{s_N})}{P(\mathcal{Q}|\hat{s_N})} = \frac{P_0(\mathcal{Q}_{split})}{P_0(\mathcal{Q})} \frac{|\mathcal{Q}|}{|\mathcal{Q}| + N} \frac{(\frac{m_{(k1)}}{m_{(k1)}+m_{(k2)}})^{-m_{(k1)}k_1} (\frac{m_{(k2)}}{m_{(k1)}+m_{(k2)}})^{-m_{(k2)}k_2}}{\binom{k_1 m_{(k1)} + k_2 m_{(k2)}}{k_1 m_{(k1)}}} \tag{5.6}$$

5.2 Cutting the $\mathcal{K}$ partition

5.2.1 1-cuts

For instance, we could compute the posterior ratio for a single splitting of the K partition; in this case we obtain, for typical cases (see D.2) and in

the limit $N \gg 1$:

$$\frac{P(\mathcal{K}_{1(12)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} \sim \Delta_0 N^{-\frac{1}{4}} \tag{5.7}$$

This effect grows with growing number of cuts.
This tells us that we can reasonably expect the posterior probabilities of partitions which contain "cuts" with respect to the $\mathcal{K}$ one (i.e. partitions that are *not coarser* then $\mathcal{K}$) to become small in the $N \gg 1$ regime.

5.3 Finer vs coarser partitions

The initial Q-expanded sum can be manipulated in order to highlight the effect of all the partitions which are *not coarser* then $\mathcal{K}$, i.e. not obtainable by $\mathcal{K}$ just by merging some of its sets. If we define:

$$\mu^{\mathcal{Q}}_{j(s)} = \frac{\rho^{\mathcal{Q}_c}_{j(s)}}{m^{\mathcal{Q}_c}_{j(s)}} \tag{5.8}$$

We can obtain (see D.3):

$$P(s|\hat{s_N}) = \sum_{\mathcal{Q}_c > \mathcal{K}} \mu^{\mathcal{Q}}_{j(s)} P(\mathcal{Q}_c|\hat{s}_N) + \mu^{\mathcal{K}}_{j(s)} P(\mathcal{K}|\hat{s}_N) \left(1 + \sum_{\mathcal{Q}_f < \mathcal{K}} \frac{\mu^{\mathcal{Q}_f}_{j(s)}}{\mu^{\mathcal{K}}_{j(s)}} \frac{P(\mathcal{Q}_f|\hat{s}_N)}{P(\mathcal{K}|\hat{s}_N)} \right) \tag{5.9}$$

Here, inside the parentheses, the "μ/μ" ratio is always less than 1 by construction, and the second ratio is a cut posterior ratio of the form studied before. We can thus conclude safely that the last parentheses can be written as $(1 + \varepsilon(N, \hat{s}_N, s))$ where ε goes to 0 as N grows. If we insert this whole posterior in 3.1 and then this latter into 3.25, we eventually get (see D.4):

$$p(s|\hat{g}) = \frac{e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j(s))} \left(1 + \eta(s)\right)}{\sum_j^{\mathcal{K}} \left[|K_j| e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)} \left(1 + \eta(s)\right) \right]} \tag{5.10}$$

Where:

$$\eta(s) = \frac{\sum_{\mathcal{Q}_f < \mathcal{K}} \mu^{\mathcal{Q}_f}_{j_{\mathcal{Q}_f}(s)} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu^{\mathcal{Q}_c}_{j_{\mathcal{Q}_c}(s)} P(\mathcal{Q}_c|\hat{s}_N)} \tag{5.11}$$

5.3.1 Perturbative treatment

We can treat $\eta(s)$ as a random variable. Then we can expand our $p(s|\hat{g})$:

$$\begin{aligned} p_\eta(s|\hat{g}) &= \frac{e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j(s))}\,(1+\eta(s))}{\sum_j^{\mathcal{K}}\left[|K_j| e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)}\left(1+\frac{1}{|K_j|}\sum_{s\in Q_j}\eta(s)\right)\right]} \\ &= \frac{e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j(s))}\,(1+\eta(s))}{\sum_j^{\mathcal{K}}\left[|K_j| e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)}\right]} \frac{\sum_j^{\mathcal{K}}\left[|K_j| e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)}\right]}{\sum_j^{\mathcal{K}}\left[|K_j| e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)}\left(1+\frac{1}{|K_j|}\sum_{s\in Q_j}\eta(s)\right)\right]} \\ &= p_0(s|\hat{g})\frac{1+\eta(s)}{1+<\eta>_{\hat{g},\eta=0}} \end{aligned} \tag{5.12}$$

So that the log-likelihood gradient becomes:

$$\begin{aligned} \partial_\lambda \log(p_\eta(\hat{s}_N|\hat{g})) =& \partial_\lambda \log(p_0(\hat{s}_N|\hat{g})) - \partial_\lambda \log\left(1+<\eta>_{\hat{g},\eta=0}\right) \\ &= N\left[\bar{\psi}^\lambda - <\psi^\lambda>_{\hat{g},\eta=0} + \frac{<\eta><\psi^\lambda> - <\eta\psi^\lambda>}{1+<\eta>}\right] \\ \approx& N\left(\bar{\psi}^\lambda - <\psi^\lambda>_{\hat{g},\eta=0} - Cov_{\vec{g},\eta=0}\left[\eta,\psi^\lambda\right]\right) \end{aligned} \tag{5.13}$$

We are interested in how the presence of $\eta(s)$ affects the definition of our sufficient statistics. $\eta(s)$, as a function of the states, can be decomposed on the ψ basis:

$$\eta(s) = \sum_\eta c_\eta \psi^\eta(j(s)) + \sum_\rho c_\rho \psi^\rho(s) \tag{5.14}$$

Where the index ρ denotes the $2^n - |\mathcal{K}|$ directions in $\vec{g}$ space along which, in the unperturbed case, we would have $\tilde{g}_\eta = 0$.
The first of these two sums does not affect the definition of the sufficient statistics, as it depends on the states s only through their frequencies; the second sum could instead add additional degrees of freedom to our model, along directions previously discarded.
If we evaluate the log-likelihood gradient along an "irrelevant" direction ρ at the unperturbed maximum ($g^\rho = 0$) we have that

$$\bar{\psi}^\rho - <\psi^\rho>_{\hat{g},\eta=0} = 0 \tag{5.15}$$

and thus:

$$\partial_\rho \log(p_\eta(\hat{s}_N|\hat{g})) = -N Cov_{\vec{g},\eta=0}\left[\eta,\psi^\lambda\right] \tag{5.16}$$

which is small if η is small. We can thus espand it at first order in the "irrelevant" directions g^ρ:

$$\partial_\rho \log(p_\eta(\hat{s}_N|\hat{g})) \approx -N g_\rho \cdot \partial_\rho Cov_{\vec{g},\eta=0}\left[\eta, \psi^\lambda\right]\Big|_{g^{\text{irr}}=0} \tag{5.17}$$

which is equivalent of having redefined the scaled log-likelihood $\mathcal{L}$:

$$\widetilde{\mathcal{L}} = \mathcal{L} - \sum_\rho^{\text{irr}} \frac{g_\rho^2}{2}\left[\partial_\rho Cov_{\vec{g},\eta=0}\left[\eta, \psi^\lambda\right]\Big|_{g^{\text{irr}}=0}\right] = \mathcal{L} - \sum_\rho^{\text{irr}} g_\rho^2 C_\rho \tag{5.18}$$

his having the form of a L2 regularizator for "irrelevant" directions. To recap: whenever the $\eta \ll 1$ assumption is justified, *we have a Bayesian justification for the use of a L2 regularizator, at least for what concerns the "irrelevant" directions.*

We can now try and characterize numerically the typical behavior of η, to check whether in practice it really is small enough to justify such perturbative treatment.

5.4 Numerical characterization of η

An exact numerical computation of $\eta(s)$ would require that we computed posterior probabilities $P(\mathcal{Q}|\hat{s}_N)$ for *all* possible partitions $\mathcal{Q}$. This is unfeasible even in simple cases due to the high number of partitions involved.
We thus adopted a twofold strategy:

- First, following the heuristic principles discussed in [25], we restricted our analysis to partitions *not clamping together* states which are not seen a similar number of times.
  This has been done by representing the set of states as an array, in which states are ordered by their frequency $\hat{k}_s$:

$s_{(1)}$	$s_{(2)}$	$s_{(3)}$	$s_{(4)}$	$s_{(5)}$	$s_{(6)}$	$s_{(7)}$	$\cdots$	$s_{(\lvert\mathcal{S}\rvert)}$

  $$\hat{k}_{s_{(1)}} \geq \hat{k}_{s_{(2)}} \geq \hat{k}_{s_{(3)}} \geq \cdots \hat{k}_{s_{(|\mathcal{S}|)}}$$

  and then generating partitions $\mathcal{Q}$ as *specific choices of sets of separators between states*:

$s_{(1)}$	$s_{(2)}$	$s_{(3)}$	$s_{(4)}$	$s_{(5)}$	$s_{(6)}$	$s_{(7)}$	$\cdots$	$s_{(\lvert\mathcal{S}\rvert)}$

  This way, we "only" have 2^{2^n-1} partitions to handle.

- Second, we resorted to a Markov Chain Monte Carlo (Metropolis-Hastings) algorithm to sample the posterior distribution $P(\mathcal{Q}|\hat{s}_N)$ on this reduced set of partitions. As a benchmark, in the low-dimensional case $n = 4$ (where we have $2^{2^4-1} = 32768$ models) we also computed the posterior exactly via exhaustive enumeration of the admissible partitions.

5.4.1 $n = 4$: Exact treatment

We generated synthetic samples from 4-spins Hamiltonians, with models chosen at random. We split the set of possible partitions into 4 categories:

- *Coarse-grained* partitions are the one obtainable from the frequency one $\mathcal{K}$ by only merging together sets of the latter (i.e. only by removal of separators);
- Q_0-*intact* partitions are those in which some set of states observed with the same frequency $\hat{k}_s \neq 0$ has been split, but the set of unobserved states has not;
- Q_{obs}-*intact* partitions are the ones in which the set of unobserved states has been split, but no observed one has;
- *Both split* partitions are the ones in which both observed and unobserved frequency sets have been split.

We plotted the marginal posterior probabilities for these four classes of partitions in figure 5.1, as a function of the sample size.
The plots are quite noisy; this is because of the fact that sampling noise here affects the very definition of the frequency partititon $\mathcal{K}_{\hat{s}_N}$ for all incremental steps, while the four classes of partitions we consider are defined *in terms* of this "noisy-defined" $\mathcal{K}_{\hat{s}_N}$. Furthermore, in these plots we are *redrawing* from the same Hamiltonian an entire brand new sample $\hat{s}_N$ for each N.
A fix is that of re-doing the simulation by subsequently adding observations to a same list of states, a little chunk ΔN at a time; then, repeat the same procedure from scratch for a certain number of other realizations of the sampling process, and finally average the results. An average over 50 realizations of the sampling process can be seen in figure 5.2.

A few comments:

- There is good agreement between the exact and approximated plots; we thus feel confident that Monte Carlo simulations applied also in higher dimensional cases will give meaningful results.
- The results in this $n = 4$ case, for our purposes, are discouraging. We see that in order to have, for instance, $P_{\text{coarse grained}} > 0.9$, we need $N \sim 5000$, which is really high for a 4-spin system (in proportion, the

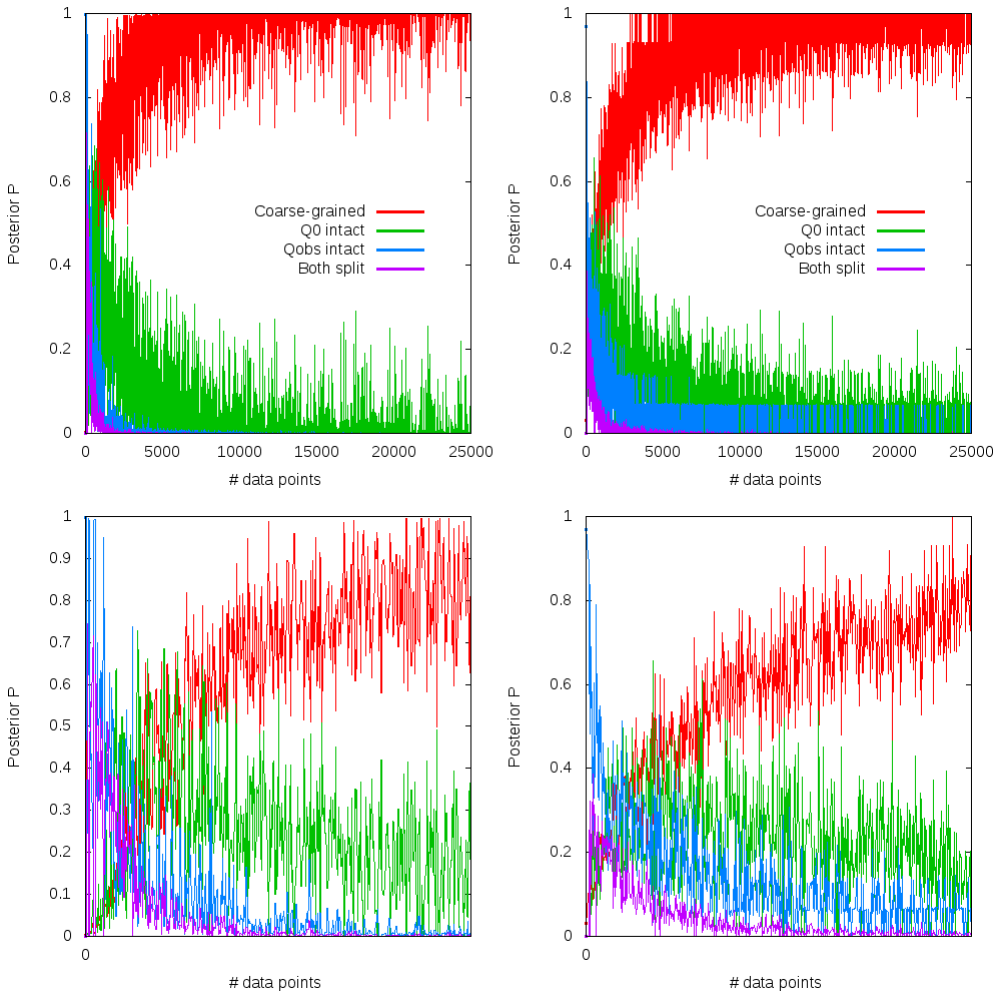

Figure 5.1: *Posterior probabilities of different types of partitions.*
Hamiltonian: $\mathcal{H} = \phi^{(1)} + \phi^{(4)} - 1.2 \cdot \phi^{(2,3,4)} - 1.3 \cdot \phi^{(1,3)}$
Top left: exact posterior probabilities $P(\mathcal{Q}|\hat{s}_N)$ computed for all 32768 admissible partitions.
Top right: MCMC estimates for the same posteriors. Each point on the plot is the result of a Metropolis-Hastings algorithm (10^5 iterations).
Bottom: Same plots, zoomed on samples $N < 3000$.
Different colors correspond to different marginal probabilities for sets of partitions of the same type (see main text).

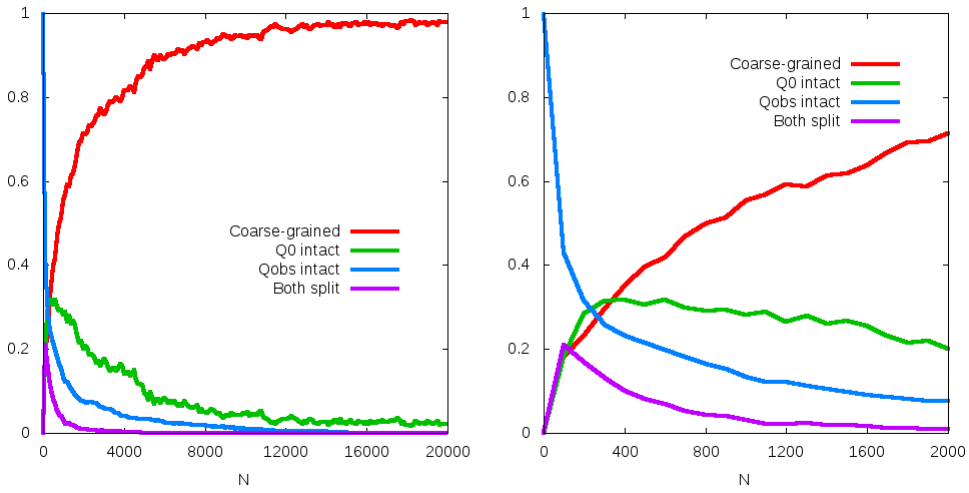

Figure 5.2: $n = 4$ *averaged posteriors*
Monte carlo estimated posterior probabilities $P(\mathcal{Q}|\hat{s}_N)$ plotted against the size of the sample set. Left: $N \leq 20000$, averaged over 50 realizations; resolution $\Delta N = 5$. Right: $N \leq 2000$, averaged over 50 realizations, resolution $\Delta N = 100$. For $N \sim 400$ the coarse grained class becomes the most probable one.

datasets used in chapter 3 were $N \sim 900$ for $n = 9$ spins!). It then seems that our perturbative approach would only be meaningful in regimes which are very far from the practical ones.

- While probabilities of splittings of the unobserved set fall rapidly to negligible values, splittings of the observed ones are very slow to be ruled out.

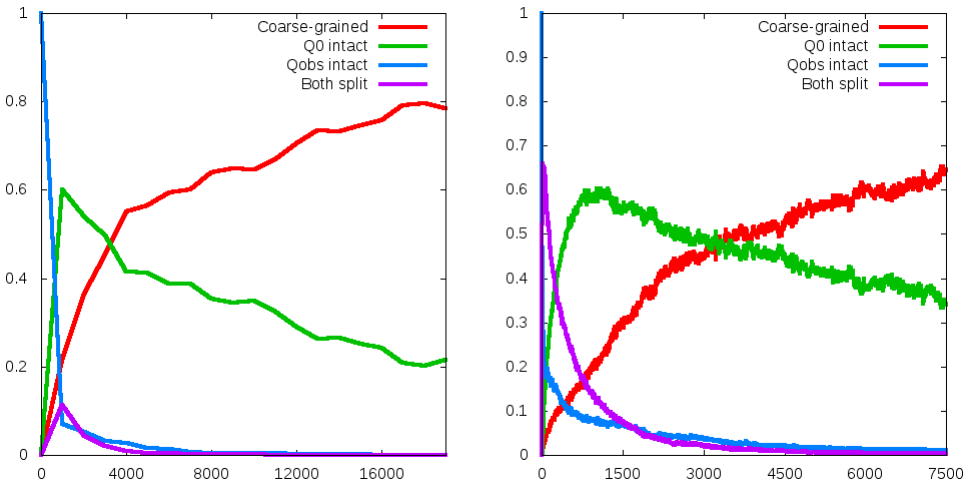

Figure 5.3: $n = 5$ *model posteriors.*
Left: $N < 20000$ (for comparison: it's the same range as the plots made for the $n = 4$ case. Average over 50 realizations, resolution $\Delta N = 1000$.
Zoom on the first $N = 7500$ sample sizes. For $N \sim 4000$ the coarse grained class becomes the most probable one. Average over 50 realizations, resolution $\Delta N = 5$.

5.4.2 $n = 5$

Results for the $n = 5$ and $n = 6$ case are plotted in figures 5.3 and 5.4. This time we just resort to Monte Carlo, since the admissible models are too many ($\sim 10^9$, $\sim 10^{19}$ respectively) for exact enumeration.

We see clearly that our situation worsens quickly with the increasing dimensionality of the system.

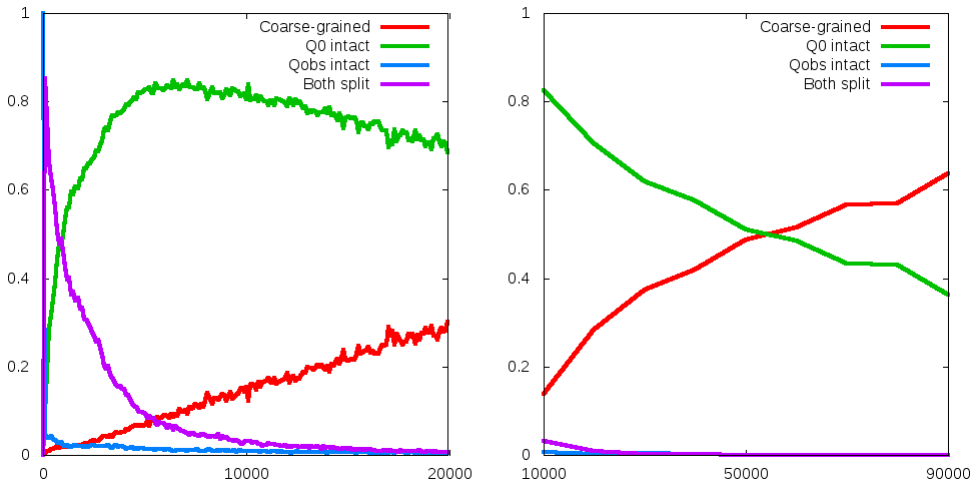

Figure 5.4: $n = 6$ *model posteriors.*
Left: $N < 20000$ (for comparison: it's the same range as the plots made for the $n = 4$ case. Average over 50 realizations, resolution $\Delta N = 100$.
Right: for $N \sim 50000$ the coarse grained class becomes the most probable one. Average over 20 realizations, resolution $\Delta N = 10000$.

5.4.3 Conclusions

We characterized precisely the typical behavior of the marginal posteriors for our different classes of models.

What we see is that the regime, in terms of sample size, in which a perturbative treatment of the ratio η would be justified is well beyond practical affordability.

The non negligibility of the "refined" partitions even for large datasets can a posteriori be attributed to the fact that in the undersampling regime (in which we are both in the $n = 5$ and $n = 6$ simulations: we never got close to $N \sim 2^{|\mathcal{S}|}$) we tend to have in samples a clamping of low-frequency states which does not correspond to simmetries in the generative distribution. In this respect a macroscopic value of the "refined" posterior signals that Bayesian analysis intrinsically aknowledges this phenomenon.

This results makes it evident that our $\mathcal{Q}$-expansion cannot, *in practice*, serve as a justification for a L2 regularization scheme.

Chapter 6

Conclusions

In the present work we described a Bayesian model selection procedure for spin models with interactions of arbitrary order. These are in astronomical numbers, even for modest n. The choice of restricting to the class of pairwise models, as is commonly done in practical applications, significantly reduces the dimesionality of the problem; yet this choice, as we have seen, cannot be justified in principle by means of Bayesian model selection arguments. Instead, we show how to peform selection in the class of mixtures, and then project the result in the space of spin models. This approach is able to spot symmetries between states that have been observed a similar number of times; these simmetries are then mapped into constraints between the spin parameters g^μ. In very simple cases, as we have checked, this mapping alone allows to retrieve the correct model - or a very small set of admissible models - in that it fixes values $g^\mu = 0$ for all those interactions ϕ^μ which are not consistent with the observed simmetries. What is left in the spin representation, after we enforce all the constraints generated by these symmetries, is a set of *sufficient statistics* ψ^η whose empirical averages allow us to compute the maximum likelihood parameters $\hat{g}^\eta$ of the model. These functions ψ^η are thus *relevant variables* for the system under study, emerging from a Bayesian procedure aimed at optimally reducing overfitting. We show that the number of different ψ^η is determined by the number of sets composing the selected partition (be it the frequency one $\mathcal{K}$, or the optimal one $\mathcal{Q}^*$) and thus it is controlled by the different frequencies observed in the data, rather than by the total number of states 2^n. This is convenient for situations in which one is in the undersampling regime, since in this case tipically $|\mathcal{K}| \ll 2^n$.

The main contribution of the present work is that of having characterized in detail the structure of the mapping between mixture and spin representation. This mapping is encoded in a matrix χ: we have shown how the information relative to different sets Q_j of the selected partition is organized in this matrix, by establishing a correspondence between these

sets and the singular values Λ_η of χ. This allowed us to understand how different sufficient statistics ψ^η correlate with different sets Q_j. We also found some identities allowing to reconstruct analitically the functions ψ^η by only knowing the singular values Λ_η and the cardinalities $|Q_j|$, a result which reduces the computational cost of retrieval of these functions.

We also explored how a fully Bayesian approach to mixture selection, via "$\mathcal{Q}$-expansion", would affect the definition of the sufficient statistics. We considered the posterior probability ratio between partitions which are not coarser than the frequency one $\mathcal{K}$, and ones that are; we showed that if this ratio becomes very small we can derive a "Bayesian regularizator" for parameter inference. We then checked numerically the typical behavior of this ratio. We found that, unluckily, it is not small for samples of reasonable size.

There are many perspectives for future investigations:

- The result that *sufficient statistics $\psi^\eta(s)$ depend on states only through their empirical frequencies* is an insightful one. It calls for further characterization of the "dynamics" of the $\mathcal{K}$ partition for sampling processes corresponding to different classes of systems; insight in this direction can probabily be gained from some recent results about "most informative samples" (see [31], [25], [32]).

- The revealed structure of the χ mapping suggests that a possibility for regularization could be that of *discarding the largest singular value* $\Lambda_{\max}$ in its spectrum: this because $\Lambda_{\max}$ turned out to correlate with the *largest* set of the selected partition, and this set, in the undersampling regime, almost always coincides with the set of unobserved states, which in turn are the ones causing the divergencies. Yet the matter is subtle, since this strategy would amount to arbitrarily discarding potentially valuable informations, in contradiction with the "Bayesian philosophy" about inference. This is an interesting path that will be followed in the near future;

- The observation made in last section of Chapter 4, that partitions seem to have special links to *loop structures* in the spin representations, calls for closer investigations. In particular, full characterization of the mapping between the two could possibly lead to an analytical recipe to reconstruct the mapping U between spin operators ϕ and sufficient statistics ψ - in a fashion similar to the discussion in Chapter 4 leading to full characterization of the W matrix (giving the values of ψ as functions of Q-states). This would arguably reduce the numerical cost of the inverse mapping $\hat{g}^\eta \to g^\mu$.

- As has been noted in [33], the exponential family of probability distributions is equipped with a hierarchical structure, by which low order

and high order interactions are entangled in non-trivial ways. It seems necessary to understand how much the emergence of spurious interactions (e.g. the four-body and eight-body ones for the "synthetic" dataset in Chapter 3) is linked to this particular structure; and to understand how this picture fits with the observation that the structure of dependencies between spin operators is preserved under order-mixing gauge transformations.

- In non-trivial cases, our inferred models will be sparse in the ψ representation, but not in the ϕ one. Extension to system with high dimensionality requires that efficient strategies for sparsification in the g^{μ} space are devised (a recipe that could be adapted is maybe the one in [34]). . Heuristics as well as more standard methods for doing so are being presently evaluated and will be the focus for further research.

Inference problems of all kinds are nowadays addressed with astonishing results by many recent devices of statistical learning theory. For instance, deep learning has changed the way we address inverse problems in physics and many other fields. Yet, efforts concerning more fundamental aspects of the kind discussed here is valuable: developing a deeper understanding of the process of separation between relevant variables, in the form of sufficient statistics, and irrelevant ones is a necessary step towards what could become, in some sense, *comprehension* of the systems under study, in the spirit of the quote with which we opened this dissertation.

Appendix A: the "a" parameter for symmetric Dirichlet prior distributions

How do we fix a value for a? The most common choices in practice are:

- a=1; this amounts to a *uniform* distribution on the space of parameters (which constitutes a *probability simplex*; in the two-states case, this corresponds to the line $(0,1) \to (1,0)$ in the $P(1), P(-1)$ plane);
- $a = \frac{1}{2}$: this turns out to be Jeffreys' reparametrization-invariant prior for Dirichlet distributions (as explicitly shown at the end of this Appendix).
- $a \to 0$. This improper prior is used in cases in which we are not even sure of what precisely is the *space of states* of our system; usage of this parameter in the prior entails that, said colorfully, *we do not even concieve things we don't observe*; the posterior probability of any unobserved state under this prior choice will be exactly equal to 0, and the alphabet will be effectively enlarged by one state any time we see a particular one that we had never previously observed.

See [35], [7] for a detailed discussion of these priors. In general, the a parameter modulates a symmetric Dirichlet distribution in a way such that a high ($\gg 1$) value of it expresses a strong prior confidence in the probabilities ρ_j being close to each other, i.e. on the distribution of states being close to uniform.

"Explorer"'s prior choice $a \to 0$: example

Let's repeat the calculation done in 3.1.2 for model selection over a 2-states system, but choosing "$a \to 0$" for our prior. We recall that the asymptotic behavior of the gamma function is such that:

$$\Gamma(\varepsilon) \sim \frac{1}{\varepsilon} \tag{6.1}$$

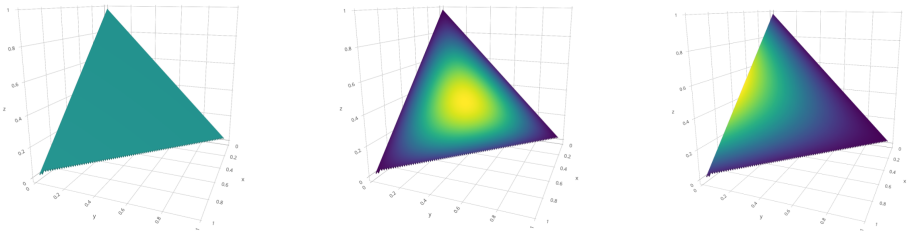

Figure 6.1: *The 3-Dirichlet distribution.*
The distribution is plotted on the 3-simplex for different values of the parameters.
Left: $(a_1 = 1, a_2 = 1, a_3 = 1)$;
Center: $(a_1 = 2, a_2 = 2, a_3 = 2)$;
Right: $(a_1 = 2, a_2 = 1, a_3 = 2)$;

This means that our posterior ratio becomes:

$$\frac{P(\mathcal{M}_1|\hat{s}_N)}{P(\mathcal{M}_0|\hat{s}_N)} \sim a \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} \frac{(\hat{k}_1 - 1)!(N - \hat{k}_1 - 1)!}{2^{1-N}(N-1)!} \longrightarrow 0, \quad \forall \hat{k}_1 \leq N \tag{6.2}$$

In this case we see that we end up selecting $\mathcal{M}_0$ irrespective of the observations made. This is clearly an unwanted behavior; in Chapter 5 we checked that the same thing happens, for $a \to 0$, also in the general case (arbitrary number of states). This is the reason why we mainly resort to the choice $a = 1$ for mixture model selection.

Why Dir($\frac{1}{2}$) is Jeffrey's prior

Assume we have a Dirichlet prior which is "s-symmetric":

$$D_a(\vec{\rho}^{(\mathcal{Q})}) = \frac{\Gamma(aQ)}{\Gamma(a)^Q} \prod_{j=1}^{Q} \left(\vec{\rho}_j^{(\mathcal{Q})}\right)^{a-1} \delta\left(\sum_{j=1}^{Q} m_j \vec{\rho}_j^{(\mathcal{Q})} - 1\right)$$

and not "j-symmetric", which would mean:

$$\tilde{D}_a(\vec{\rho}^{(\mathcal{Q})}) = \frac{\Gamma(aQ)}{\Gamma(a)^Q} \prod_{j=1}^{Q} \left(\vec{\rho}_j^{(\mathcal{Q})}\right)^{a-1} \delta\left(\sum_{j=1}^{Q} \vec{\rho}_j^{(\mathcal{Q})} - 1\right)$$

Now, it is easy to show that the Fisher Information matrix of this first distribution is:

$$J_{ij} = \frac{\delta_{ij}}{p_i} + \frac{1}{p_Q} \quad \forall i, j \in [1, Q-1]$$

One can prove that the determinant of a matrix obtained by summing a $N \times N$ constant (b) matrix to a diagonal ($\{d_i\}_i$) one is:

$$\det A = \prod_{i=1}^{N} \left(d_i - b\right) + b \sum_{i=1}^{N} \prod_{j \neq i} \left(d_i - b\right)$$

So that:

$$\begin{aligned}
\det J &= \prod_{i=1}^{Q-1}\left(\frac{1}{p_i}+\frac{1}{p_Q}-\frac{1}{p_Q}\right)+\frac{1}{p_Q}\sum_{i=1}^{Q-1}\prod_{j\neq i}\frac{1}{p_j}= \\
&= \left(\prod_{i=1}^{Q-1}\frac{1}{p_i}\right)\left(1+\frac{1}{p_Q}\sum_{i=1}^{Q-1}p_i\right)= \\
&= \left(\prod_{i=1}^{Q-1}\frac{1}{p_i}\right)\left(1+\frac{1-p_Q}{p_Q}\right)= \\
&= \left(\prod_{i=1}^{Q-1}\frac{1}{p_i}\right)\frac{1}{p_Q}= \\
&= \prod_{i=1}^{Q}\frac{1}{p_i}
\end{aligned} \tag{6.3}$$

The Jeffreys' volume element is $\sqrt{\det J(\vec{\rho}^{(Q)})}$.
This leads immediately to the choice of $a = \frac{1}{2}$.

Appendix B: $\vec{g}$ estimation via log-likelihood maximization

As anticipated, the recipe we will mainly use for simulations is that of searching for the parameter values that maximize the logligelihood of the data. We recall that this loglikelihood reads:

$$\begin{aligned}\log P(\hat{s}_N|\tilde{\mathbf{g}}) =& \sum_{\eta>0}^{\mathcal{M}} \sum_{i}^{N} g^\eta \psi^\eta(\mathbf{s}^{(i)}) - N \log \mathcal{Z}_{\mathcal{M}}(\tilde{\mathbf{g}}) \\ =& N \left(\tilde{g}^\eta \overline{\psi^\eta} - \log \mathcal{Z}_{\mathcal{M}}(\tilde{\mathbf{g}}) \right)\end{aligned} \tag{6.4}$$

where we defined the *empirical averages* $\overline{\psi^\eta} = \frac{1}{N} \sum_i^N \psi^\eta(\mathbf{s}^{(i)})$. Maximum of this function is found imposing:

$$\begin{aligned}0 =& \frac{\partial}{\partial g^\eta} \log P(\hat{s}_N|\tilde{\mathbf{g}}) = N \left(\overline{\psi^\eta} - \partial_\eta \log \mathcal{Z}_{\mathcal{M}}(\tilde{\mathbf{g}}) \right) \\ =& N \left(\overline{\psi^\eta} - < \psi^\eta >_{P(s|g)} \right)\end{aligned} \tag{6.5}$$

So that likelihood is maximized for those values of the couplings such that the *empirical averages of ψ functions coincide with their ensemble average.* In other words, the maximum likelihood parameters $\tilde{\mathbf{g}}^*$ reproduce exactly the statistics of the sample; in this respect, it is easy to check that the probability of any state $\mathbf{s}$ under this choice of parameters will coincide with the observed frequency of that state:

$$P(\mathbf{s}|\mathbf{g}^*) \equiv \frac{\hat{k}_s}{N} \tag{6.6}$$

From this last relation we can convince ourselves that as long as we don't arbitrarily prune our model by eliminating some operators, there's no point in numerically performing the maximization of loglikelihood - we can obtain directly the components of $\mathbf{g}^*$ via formula 3.1.

Appendix C: one more identity for the spectrum of χ

Once we verify that the sufficient statistics are orthogonal:

$$\begin{aligned}\sum_s \left[\frac{\psi^\eta(s)\psi^\rho(s)}{2^n}\right] &= \frac{1}{2^n}\sum_s \left[\sum_{\mu>0}\sum_{\nu>0} U_{\mu\eta}U_{\nu\rho}\phi^\mu(s)\phi^\nu(s)\right] \\ &= \sum_{\mu>0}\sum_{\nu>0} U_{\mu\eta}U_{\nu\rho}\delta_{\mu\nu} \\ &= \sum_{\mu>0} U_{\mu\eta}U_{\mu\rho} = \delta_{\eta\rho} \quad \forall\eta,\rho : \Lambda_\eta \neq 0, \Lambda_\rho \neq 0\end{aligned} \tag{6.7}$$

we can compute, using (4.2):

$$\begin{aligned}\sum_s \left[\frac{\psi^\eta(s)\psi^\rho(s)}{2^n}\right] = \delta_{\eta\rho} &= \frac{1}{\sqrt{\lambda_\eta}}\frac{1}{\sqrt{\lambda_\rho}}\sum_j \left\{\omega_j\left[(W_{\eta j}-\overline{W_\eta})(W_{\rho j}-\overline{W_\rho})\right]\right\} \\ &= \frac{\overline{W_\eta W_\rho}-\overline{W}_\eta\overline{W}_\rho}{\sqrt{\lambda_\eta\lambda_\rho}}\end{aligned} \tag{6.8}$$

And thus:

$$\overline{W_\eta W_\rho} = \overline{W}_\eta\overline{W}_\rho \quad \eta \neq \rho \tag{6.9}$$

$$\lambda_\eta = \overline{W_\eta^2} - \overline{W_\eta}^2 \tag{6.10}$$

Now summing over η, and using the orthogonality of W:

$$\begin{aligned}
\sum_{\eta} \lambda_{\eta} &= \sum_{\eta} \sum_{j} W_{\eta j}^2 \omega_j - \sum_{\eta} \sum_{j} \sum_{k} \omega_j \omega_k W_{\eta j} W_{\eta k} \\
&= \sum_{j} \omega_j \cdot 1 - \sum_{\eta} \overline{W_{\eta}}^2 \\
&= 1 - \sum_{\eta} \overline{W_{\eta}}^2 \\
&= 1 - \sum_{j} \sum_{k} \omega_j \omega_k \delta_{jk} \\
&= 1 - \sum_{j} \omega_j^2
\end{aligned}$$

meaning that:

$$\sum_{j} \omega_j^2 \equiv \sum_{\eta} \overline{W_{\eta}}^2 \tag{6.11}$$

$$\sum_{\eta} \left(\lambda_{\eta} + \omega_{\eta}^2\right) = \sum_{\eta} \left(\lambda_{\eta} + \overline{W_{\eta}}^2\right) = 1 \tag{6.12}$$

Which we can manipulate:

$$\sum_{\eta} \left(\lambda_{\eta} + \omega_{\eta}^2\right) = \sum_{\eta_{nd}} \lambda_{\eta_{nd}} + \sum_{\eta_{deg}} \lambda_{\eta_{deg}} + \sum_{j} \omega_j^2 = 1$$

$$\sum_{\eta_{nd}} \lambda_{\eta_{nd}} = 1 - \sum_{j} \omega_j^2 - \sum_{\eta_{deg}} \lambda_{\eta_{deg}}$$

And, finally:

$$\sum_{\eta_{nd}} \lambda_{\eta_{nd}} = 1 - \sum_{j} \omega_j^2 - \sum_{\eta_{deg}} \omega_{\eta} \tag{6.13}$$

Appendix D: Detail of calculations for the $|\mathcal{Q}|$-expansion

D.0: State posterior probabilities $P(s|\hat{s}_N, \mathcal{Q})$

Although we already stated precisely what this term is equal to, let's get rid of any possible doubt by a raw calculation:

$$\begin{aligned}
P(s|\hat{s}_N, \mathcal{Q}) &= \int_{\vec{\rho}_\mathcal{Q}} d\vec{\rho_\mathcal{Q}} P(s|\vec{\rho}_\mathcal{Q}, \mathcal{Q}, \hat{s}_N) P(\vec{\rho}_\mathcal{Q}|\mathcal{Q}, \hat{s}_N) \\
&= \int_{\vec{\rho}_\mathcal{Q}} d\vec{\rho_\mathcal{Q}} \frac{\rho_{\mathcal{Q}j(s)}}{m_{\mathcal{Q}j(s)}} P(\vec{\rho}_\mathcal{Q}|\mathcal{Q}, \hat{s}_N) \\
&= \int_{\vec{\rho}_\mathcal{Q}} d\vec{\rho_\mathcal{Q}} \frac{\rho_{\mathcal{Q}j(s)}}{m_{\mathcal{Q}j(s)}} P(\hat{s}_N|\vec{\rho}_\mathcal{Q}, \mathcal{Q}) \frac{P(\vec{\rho}_\mathcal{Q}|\mathcal{Q})}{P(\hat{s}_N|\mathcal{Q})} \\
&= \frac{1}{P(\hat{s}_N|\mathcal{Q})} \int_{\vec{\rho}_\mathcal{Q}} d\vec{\rho_\mathcal{Q}} \frac{\rho_{\mathcal{Q}j(s)}}{m_{\mathcal{Q}j(s)}} \prod_{j=1}^{q} \left[\frac{{\rho_j}^{K_{\mathcal{Q}j}}}{m_{\mathcal{Q}j}^{K_{\mathcal{Q}j}}} \right] P(\vec{\rho}_\mathcal{Q}|\mathcal{Q}) \\
&= \frac{1}{P(\hat{s}_N|\mathcal{Q})} \int_{\vec{\rho}_\mathcal{Q}} d\vec{\rho_\mathcal{Q}} \frac{\rho_{\mathcal{Q}j(s)}}{m_{\mathcal{Q}j(s)}} \prod_{j=1}^{q} \left[\frac{{\rho_j}^{K_{\mathcal{Q}j}+a_\mathcal{Q}-1}}{m_{\mathcal{Q}j}^{K_{\mathcal{Q}j}}} \right] \frac{\Gamma(a_Q Q)}{\Gamma(a_\mathcal{Q})^Q} \delta(\sum_j \rho_j - 1) \\
&= \left\{ \frac{1}{P(\hat{s}_N|\mathcal{Q})} \frac{\Gamma(a_Q Q)}{\Gamma(a_\mathcal{Q})^Q \Gamma(a_\mathcal{Q} Q + N)} \prod_{j=1}^{Q} \left[\frac{\Gamma(K_{\mathcal{Q}j} + a_\mathcal{Q})}{m_{\mathcal{Q}j}^{K_{\mathcal{Q}j}}} \right] \right\} \frac{1}{m_{\mathcal{Q}j}} \frac{K_{\mathcal{Q}j(s)} + a_\mathcal{Q}}{N + a_\mathcal{Q}\mathcal{Q}}
\end{aligned} \tag{6.14}$$

Now, if I sum this over s I get, enforcing normalization:

$$P(\hat{s}_N|\mathcal{Q}) = \frac{\Gamma(a_Q Q)}{\Gamma(a_\mathcal{Q})^Q \Gamma(a_\mathcal{Q} Q + N)} \prod_{j=1}^{Q} \left[\frac{\Gamma(K_{\mathcal{Q}j} + a_\mathcal{Q})}{m_{\mathcal{Q}j}^{K_{\mathcal{Q}j}}} \right] \tag{6.15}$$

So that, finally:

$$P(s|\hat{s}_N, \mathcal{Q}) = \frac{1}{m_{\mathcal{Q}j(s)}} \frac{K_{\mathcal{Q}j(s)} + a_\mathcal{Q}}{N + a_\mathcal{Q}\mathcal{Q}} \tag{6.16}$$

Which is an instance of Laplace's rule of succession for model $\mathcal{Q}$.

D.1: The case "$a \to 0$"

Maximum likelihood parameter estimation is a popular strategy, and we have seen it is equivalent to the $a \to 0$ estimator for $P(s|\hat{s}_N)$. It would be tempting to try and transpose that choice of prior parameter in this new setting by letting all $a_{\mathcal{Q}} \to 0$, but there are important issues with this operation.
First, let's see what happens when $a_{\mathcal{Q}} = a \ \forall \mathcal{Q}$, for a single-cut posterior ratio (starting from any partition $\mathcal{Q}$). Recall:

$$\frac{P(\mathcal{Q}_{1(12)}|\hat{s_N})}{P(\mathcal{Q}|\hat{s_N})} = \frac{P_0(\mathcal{Q}_{1(12)})}{P_0(\mathcal{Q})} \frac{\Gamma(aQ_m + a)\Gamma(aQ_m + N)\Gamma(a + K_1)\Gamma(a + K_2)}{\Gamma(aQ_m)\Gamma(a)\Gamma(aQ_m + a + N)\Gamma(a + K_1 + K_2)} \eta_1^{-K_1} \eta_2^{-K_2} \tag{6.17}$$

Now if $a \to 0$:

$$\frac{P(\mathcal{Q}_{1(12)}|\hat{s_N})}{P(\mathcal{Q}|\hat{s_N})} \approx a \left\{ \frac{P_0(\mathcal{Q}_{1(12)})}{P_0(\mathcal{Q})} \frac{q_m}{q_m + 1} B(K_1, K_2) \eta_1^{-K_1} \eta_2^{-K_2} \right\} + o(a^2) \tag{6.18}$$

meaning that:

$$P(\mathcal{Q}|\hat{s}_N) \to \delta_{\mathcal{Q},Q_0} \tag{6.19}$$

where $\mathcal{Q}_0$ is the partition composed by a single set.
We thus see that if we perform this limit in this way, we get for small values of a an unwanted, strong bias in favor of *merging* sets. In the limit, this would mean that our posterior will assign equal probability to all possible outcomes, irrespective of the observations made.
This is formally equivalent to the posterior we would find if we performed inference on parameters with the finest model $\mathcal{Q} = \mathcal{S}$:

$$P(s|\hat{s}_N, I) = \int d\vec{\mu}_{\mathcal{S}} \prod_{s \in \mathcal{S}} \left[{\rho_s}^{k_s} \right] P_0(\vec{\rho}, I_a)$$

in the case of "$a \to \infty$" as a prior choice. A legitimate question would then be whether by relaxing the condition $a_{\mathcal{Q}} = a \ \forall \mathcal{Q}$ we can find a way of letting all parameters go to 0 without getting a singular posterior: it can be shown that this is possible only if we assume that:

$$a_{\mathcal{Q}} = a(|\mathcal{Q}|) = a_2^{|\mathcal{Q}|-1}$$

and then proceed with $a_2 \to 0$. In that particular case we get:

$$\frac{P(\mathcal{Q}_{split}|\hat{s_N})}{P(\mathcal{Q}|\hat{s_N})} = \frac{P_0(\mathcal{Q}_{split})}{P_0(\mathcal{Q})}\frac{Q}{Q+1}\frac{\eta_1^{-K_1}\eta_2^{-K_2}}{\binom{K_1+K_2-2}{K_1-1}}$$
$$= \frac{P_0(\mathcal{Q}_{split})}{P_0(\mathcal{Q})}\frac{Q}{Q+1}\frac{(K_1+K_2)(K_1+K_2-1)}{K_1K_2}\frac{\eta_1^{-K_1}\eta_2^{-K_2}}{\binom{K_1+K_2}{K_1}} \tag{6.20}$$

If we try and compute the $N >> 1$ asymptotics for this object as we did in the $a = 1$ case, we see that in this limit the ratio diverges (splitting becomes infinitely times more favorable), for all possible samples. It is difficult to interpret this result, and this calls for a more careful investigation of what kind of information we are really representing when we choose the values $\{a_{\mathcal{Q}}\}$.

D.2: cutting the $\mathcal{K}$ partition

1-cuts

The posterior ratio for a single splitting is:

$$\frac{P(\mathcal{K}_{1(12)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} = \frac{P_0(\mathcal{K}_{1(12)})}{P_0(\mathcal{K})}\frac{K}{K+N}\frac{(\frac{m_{k(1)}}{m_{k(1)}+m_{k(2)}})^{-km_{k(1)}}(\frac{m_{k(2)}}{m_{k(1)}+m_{k(2)}})^{-km_{k(2)}}}{\binom{k(m_{k(1)}+m_{k(2)}))}{km_{k(1)}}} \tag{6.21}$$

that if ($km_{k(1)} >> 1$, $km_{k(2)} >> 1$) becomes

$$\frac{P(\mathcal{K}_{1(12)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} \approx \frac{P_0(\mathcal{K}_{1(12)})}{P_0(\mathcal{K})}\frac{K}{K+N}\sqrt{\frac{2\pi k m_{k(1)}m_{k(2)}}{m_{k(1)}+m_{k(2)}}}\Big(1+\Delta\Big), \tag{6.22}$$

$$\Delta = \frac{1-(m_{k(1)}+m_{k(2)})}{12k(m_{k(1)}+m_{k(2)})} + o(k^{-2}) + o((km)^{-2}).$$

We can work out a decent bound for this expression in the large N limit. In typical cases, both K and km_k are of order $\sim \sqrt{N}$. Now if we realize that, in the above expression for the posterior ratio, we can rewrite:

$$\sqrt{\frac{2\pi k m_{k(1)}m_{k(2)}}{m_{k(1)}+m_{k(2)}}} = \sqrt{2\pi k(m_{k(1)}+m_{k(2)})\eta_1\eta_2} = \sqrt{2\pi k m_k \eta_1\eta_2}$$

(from now on $\eta_1 = \frac{m_{k(1)}}{m_{k(1)}+m_{k(2)}}$ and so on), then we can substitute and get:

$$\begin{aligned}\frac{P(\mathcal{K}_{1(12)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} &\sim \frac{P_0(\mathcal{K}_{1(12)})}{P_0(\mathcal{K})}\frac{K}{K+N}\sqrt{\frac{2\pi k m_{k(1)}m_{k(2)}}{m_{k(1)}+m_{k(2)}}}\Big(1+\cdots\Big) \sim \\ &\sim \frac{P_0(\mathcal{K}_{1(12)})}{P_0(\mathcal{K})}\frac{\sqrt{N}}{\sqrt{N}+N}\sqrt{\sqrt{N}}\sqrt{2\pi\eta_1\eta_2}\Big(1+\cdots\Big) \sim \\ &\sim \Delta_0 N^{-\frac{1}{4}}\end{aligned} \tag{6.23}$$

which is the result used in the text.

2-cuts

We can also compute at once the probability for many splits. Let's first exhamine the case in which we are allowed to split each $\mathcal{K}$-set only once (we'll call these *separate* splittings):

$$\begin{aligned}\frac{P(\mathcal{K}_{2s}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} =& \frac{P_0(\mathcal{K}_{2s(12)(34)})}{P_0(\mathcal{K})}\frac{K}{K+N}\frac{K+1}{K+N+1}\cdot \\ &\cdot \frac{(\eta_1)^{-k_{12}m_{k(1)}}(\eta_2)^{-k_{12}m_{k(2)}}}{\binom{k_{12}(m_{k(1)}+m_{k(2)}))}{k_{12}m_{k(1)}}} \cdot \frac{(\eta_3)^{-k_{34}m_{k(3)}}(\eta_4)^{-k_{34}m_{k(4)}}}{\binom{k_{34}(m_{k(3)}+m_{k(4)}))}{k_{34}m_{k(3)}}}\end{aligned} \tag{6.24}$$

(in which: $\eta_1 = \frac{m_{k(1)}}{m_{k(1)}+m_{k(2)}}$ and so on) leading to:

$$\begin{aligned}\frac{P(\mathcal{K}_{2s(12)(34)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} =& \frac{P_0(\mathcal{K}_{2s(12)(34)})P_0(\mathcal{K})}{P_0(\mathcal{K}_{1s(12)})P_0(\mathcal{K}_{1s(34)})}\frac{P(\mathcal{K}_{1s(12)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})}\frac{P_(\mathcal{K}_{1s(34)}|\hat{s_N})}{P_(\mathcal{K}|\hat{s_N})}\cdot \\ &\cdot \left(1+\frac{1}{K}\right)\left(1-\frac{1}{N+K+1}\right)\end{aligned} \tag{6.25}$$

which is an exact expression.

c-cuts

The probability for c separate cuts $c < (K - \sum_{j=1}^{K} \delta_{m_j,1})$ is then:

$$\frac{P(\mathcal{K}_{cs(c_1)(c_2)\ldots}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})} = \frac{P_0(\mathcal{K}_{cs(c_1)(c_2)\ldots})P_0(\mathcal{K})^{c-1}}{\prod_{j=1}^{c} P_0(\mathcal{K}_{1(c_j)})} \prod_{j=1}^{c} \left[\frac{P(\mathcal{K}_{1(c_j)}|\hat{s_N})}{P(\mathcal{K}|\hat{s_N})}\right] \cdot \cdot \prod_{j=1}^{c-1} \left[\left(1 + \frac{j}{K}\right)\left(1 - \frac{1}{N+K+j}\right)\right] \tag{6.26}$$

this also being an exact formula.

D.3: Finer vs coarser partitions

We can compute:

$$\begin{aligned}
P(s|\hat{s_N}) &= \sum_{\mathcal{Q}} P(s|\hat{s}_N, \mathcal{Q}) P(\mathcal{Q}|\hat{s}_N) \\
&= \sum_{\mathcal{Q}_c > \mathcal{K}} P(s|\hat{s}_N, \mathcal{Q}_c) P(\mathcal{Q}_c|\hat{s}_N) \\
&+ P(s|\hat{s}_N, \mathcal{K}) P(\mathcal{K}|\hat{s}_N) \left(1 + \sum_{\mathcal{Q}_f < \mathcal{K}} \frac{P(s|\hat{s}_N, \mathcal{Q}_f)}{P(s|\hat{s}_N, \mathcal{K})} \frac{P(\mathcal{Q}_f|\hat{s}_N)}{P(\mathcal{K}|\hat{s}_N)}\right) \\
&= \sum_{\mathcal{Q}_c > \mathcal{K}} P(s|\hat{s}_N, \mathcal{Q}_c) P(\mathcal{Q}_c|\hat{s}_N) \\
&+ P(s|\hat{s}_N, \mathcal{K}) P(\mathcal{K}|\hat{s}_N) \left(1 + \sum_{\mathcal{Q}_f < \mathcal{K}} \frac{m_{\mathcal{K}j(s)}}{m_{\mathcal{Q}j(s)}} \frac{K_{\mathcal{Q}j(s)} + a_{\mathcal{Q}}}{K_{\mathcal{K}j(s)} + a_{\mathcal{K}}} \frac{N + a_{\mathcal{K}}\mathcal{K}}{N + a_{\mathcal{Q}}\mathcal{Q}} \frac{P(\mathcal{Q}_f|\hat{s}_N)}{P(\mathcal{K}|\hat{s}_N)}\right) \\
&= \sum_{\mathcal{Q}_c > \mathcal{K}} \mu^{\mathcal{Q}}_{j(s)} P(\mathcal{Q}_c|\hat{s}_N) + \mu^{\mathcal{K}}_{j(s)} P(\mathcal{K}|\hat{s}_N) \left(1 + \sum_{\mathcal{Q}_f < \mathcal{K}} \frac{\mu^{\mathcal{Q}_f}_{j(s)}}{\mu^{\mathcal{K}}_{j(s)}} \frac{P(\mathcal{Q}_f|\hat{s}_N)}{P(\mathcal{K}|\hat{s}_N)}\right)
\end{aligned} \tag{6.27}$$

D.4: Perturbed maximum entropy distribution

$$
\begin{aligned}
g^{\mu} &= 2^{-n} \sum_{s} \phi^{\mu}(s) \log(P(s|\hat{s_N})) \\
&= 2^{-n} \sum_{s} \phi^{\mu}(s) \log \left(\sum_{\mathcal{Q}_c > \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N) + \mu_{j(s)}^{\mathcal{K}} P(\mathcal{K}|\hat{s}_N) \left(1 + \varepsilon(N, \hat{s}_N, s)\right) \right) \\
&= 2^{-n} \sum_{s} \phi^{\mu}(s) \log \left(\left(\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N) \right) \left(1 + \varepsilon(N, \hat{s}_N, s) \frac{\mu_{j(s)}^{\mathcal{K}} P(\mathcal{K}|\hat{s}_N)}{\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)} + o(\varepsilon^2) \right) \right. \\
&= 2^{-n} \sum_{s} \phi^{\mu}(s) \log \left(\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N) \right) + 2^{-n} \sum_{s} \phi^{\mu}(s) \log \left(1 + C(s, \hat{s}_N) \varepsilon(N, \hat{s}_N, s)\right) \\
&= 2^{-n} \sum_{j}^{\mathcal{K}} \chi_j^{\mu} \log f(j(s)) + 2^{-n} \sum_{s} \phi^{\mu}(s) \log \left(1 + C(s, \hat{s}_N) \varepsilon(N, \hat{s}_N, s)\right)
\end{aligned}
\tag{6.28}
$$

Now, if we decompose χ via SVD and evaluate the sum $\sum_{\mu} g^{\mu} \phi^{\mu}(s)$:

$$
\begin{aligned}
\sum_{\mu} g^{\mu} \phi^{\mu}(s) &= \sum_{\lambda} \tilde{g}^{\lambda} \psi^{\lambda}(s) + 2^{-n} \sum_{\mu} \phi^{\mu}(s) \sum_{r} \phi^{\mu}(r) \log \left(1 + C(r, \hat{s}_N) \varepsilon(N, \hat{s}_N, r)\right) \\
&= \sum_{\lambda} \tilde{g}^{\lambda} \psi^{\lambda}(s) + \log \left(1 + C(s, \hat{s}_N) \varepsilon(N, \hat{s}_N, s)\right)
\end{aligned}
\tag{6.29}
$$

Note that:

$$
C(s, \hat{s}_N) \varepsilon(N, \hat{s}_N, s) = \frac{\sum_{\mathcal{Q}_f < \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)} \tag{6.30}
$$

So that our final $p(s|\hat{g})$ looks like this:

$$
p(s|\hat{g}) = \frac{1}{\mathcal{Z}_{\varepsilon}} \exp \left(\sum_{\lambda} \tilde{g}^{\lambda} \psi^{\lambda}(s) + \log \left(1 + \frac{\sum_{\mathcal{Q}_f < \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)} \right) \right) \tag{6.31}
$$

In which:

$$
\mathcal{Z}_{\varepsilon} = \sum_{s} \left[e^{\sum_{\lambda} \tilde{g}^{\lambda} \psi^{\lambda}(s)} \left(1 + \frac{\sum_{\mathcal{Q}_f < \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c \geq \mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)} \right) \right] \tag{6.32}
$$

Recall now that the sufficient statistics $\psi^{\eta}(s)$ depend on the individual sets only through the index of the partition set they belong to:

$$
\psi^{\eta}(s) = \psi^{\eta}(j_{\mathcal{Q}}(s)) \tag{6.33}
$$

With this in mind, we can write:

$$\begin{aligned}\mathcal{Z}_\varepsilon &= \sum_j^{\mathcal{K}} e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)} \sum_{s\in j} \left(1 + \frac{\sum_{\mathcal{Q}_f<\mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c\geq\mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)}\right) \\ &= \sum_j^{\mathcal{K}} e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)} \left(|K_j| + \frac{\sum_{s\in j}\sum_{\mathcal{Q}_f<\mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c\geq\mathcal{K}} \mu_j^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)}\right)\end{aligned} \tag{6.34}$$

That, for simmetry, equals:

$$\mathcal{Z}_\varepsilon = \sum_j^{\mathcal{K}} e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)} |K_j| \left(1 + \frac{\sum_{\mathcal{Q}_f<\mathcal{K}} \mu_j^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c\geq\mathcal{K}} \mu_j^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)}\right) \tag{6.35}$$

so that we can write a heavy but complete expression for $p(s|\hat{g})$:

$$p(s|\hat{g}) = \frac{e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j(s))} \left(1 + \frac{\sum_{\mathcal{Q}_f<\mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c\geq\mathcal{K}} \mu_{j(s)}^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)}\right)}{\sum_j^{\mathcal{K}} \left[|K_j| e^{\sum_\lambda \tilde{g}^\lambda \tilde{\psi}^\lambda(j)} \left(1 + \frac{\sum_{\mathcal{Q}_f<\mathcal{K}} \mu_j^{\mathcal{Q}_f} P(\mathcal{Q}_f|\hat{s}_N)}{\sum_{\mathcal{Q}_c\geq\mathcal{K}} \mu_j^{\mathcal{Q}_c} P(\mathcal{Q}_c|\hat{s}_N)}\right)\right]} \tag{6.36}$$

Bibliography

[1] Charles S. Peirce. *Essays in the Philosophy of Science*. New York: Liberal Arts Press, 1957.

[2] E. Grant, P.E.E. Grant, G. Basalla, and O. Hannaway. *The Foundations of Modern Science in the Middle Ages: Their Religious, Institutional and Intellectual Contexts*. Cambridge Studies in the History of Science. Cambridge University Press, 1996.

[3] Eugene P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. richard courant lecture in mathematical sciences delivered at new york university, may 11, 1959. *Communications on Pure and Applied Mathematics*, 13(1):1–14, 1960.

[4] Alberto Beretta, Claudia Battistin, Clélia de Mulatier, Iacopo Mastromatteo, and Matteo Marsili. The Stochastic Complexity of Spin Models: Are Pairwise Models Really Simple? *Entropy*, 20:739, Sep 2018.

[5] Luigi Gresele. A Heuristic for Model Selection for Spin Models of arbitrary order. Master's thesis, Politecnico di Torino, 2017.

[6] Luigi Gresele and Matteo Marsili. On Maximum Entropy and Inference. *Entropy*, 19:642, Nov 2017.

[7] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, Cambridge, 2003.

[8] Gavin C. Cawley and Nicola L.C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107, August 2010.

[9] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

[10] Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. *Encyclopedia of Machine Learning*, 01 2011.

[11] Vijay Balasubramanian. A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions. *arXiv e-prints*, pages adap–org/9601001, Jan 1996.

[12] I J Myung, V Balasubramanian, and M A Pitt. Counting probability distributions: differential geometry and model selection. *Proc Natl Acad Sci U S A*, 97(21):11170–11175, October 2000.

[13] T. Cover and J. Thomas. *Elements of information theory*. John Wiley and Sons, Inc., 1991.

[14] S. Amari. *Differential-geometrical methods in statistics*. Lecture notes in statistics. Springer-Verlag, 1985.

[15] Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS Computational Biology*, 3:e189, Jan 2007.

[16] Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of chemical physics*, 143(1), July 2015.

[17] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66:197–261, Jul 2017.

[18] Adam A. Margolin, Kai Wang, Andrea Califano, and Ilya Nemenman. Multivariate dependence and genetic networks inference. *arXiv e-prints*, page arXiv:1001.1681, Jan 2010.

[19] T. J. Sejnowski. Higher-order Boltzmann machines. In *American Institute of Physics Conference Series*, volume 151 of *American Institute of Physics Conference Series*, pages 398–403, August 1986.

[20] E. J. G. Pitman and J. Wishart. Sufficient statistics and intrinsic accuracy. *Proceedings of the Cambridge Philosophical Society*, 32:567, 1936.

[21] Lina Merchan and Ilya Nemenman. On the sufficiency of pairwise interactions in maximum entropy models of networks. *Journal of Statistical Physics*, 162(5):1294–1308, Mar 2016.

[22] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147 – 169, 1985.

[23] Shan Yu, Hongdian Yang, Hiroyuki Nakahara, Gustavo S. Santos, Danko Nikolić, and Dietmar Plenz. Higher-order interactions characterized in cortical activity. *The Journal of Neuroscience*, 31(48):17514–17526, 2011. Exported from https://app.dimensions.ai on 2019/03/01.

[24] Claudia Battistin, Benjamin Dunn, and Yasser Roudi. Learning with unknowns: Analyzing biological data in the presence of hidden variables. *Current Opinion in Systems Biology*, 1, 01 2017.

[25] Ariel Haimovici and Matteo Marsili. Criticality of mostly informative samples: a bayesian model selection approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(10):P10013, oct 2015.

[26] Ilya Nemenman, Fariel Shafee, and William Bialek. Entropy and inference, revisited. *arXiv e-prints*, page physics/0108025, Aug 2001.

[27] R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925.

[28] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[29] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson. Large pseudocounts and L_2-norm penalties are necessary for the mean-field inference of ising and potts models. *Phys. Rev. E*, 90:012132, Jul 2014.

[30] Edward D. Lee, Chase P. Broedersz, and William Bialek. Statistical mechanics of the us supreme court. *Journal of Statistical Physics*, 160(2):275–301, Jul 2015.

[31] Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. On sampling and modeling complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2013:09003, Sep 2013.

[32] Juyong Song, Matteo Marsili, and Junghyo Jo. Resolution and Relevance Trade-offs in Deep Learning. *arXiv e-prints*, page arXiv:1710.11324, Oct 2017.

[33] S. . Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, July 2001.

[34] Nicola Bulso, Matteo Marsili, and Yasser Roudi. Sparse model selection in the highly under-sampled regime. *Journal of Statistical Mechanics: Theory and Experiment*, 9:093404, Sep 2016.

[35] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, Sep. 1968.